Improved time complexity analysis of the Simple Genetic Algorithm

Oliveto, Pietro S.; Witt, Carsten

Published in:
Theoretical Computer Science

Link to article, DOI:
10.1016/j.tcs.2015.01.002

Publication date:
2015

Document Version
Peer reviewed version

Improved Time Complexity Analysis of the Simple Genetic Algorithm☆

Pietro S. Oliveto[a,1], Carsten Witt[b]

[a]*Department of Computer Science, University of Sheffield, United Kingdom*
[b]*DTU Compute, Technical University of Denmark, Denmark*

Abstract

A runtime analysis of the Simple Genetic Algorithm (**SGA**) for the OneMax problem has recently been presented proving that the algorithm with population size $\mu \leq n^{1/8-\varepsilon}$ requires exponential time with overwhelming probability. This paper presents an improved analysis which overcomes some limitations of the previous one. Firstly, the new result holds for population sizes up to $\mu \leq n^{1/4-\varepsilon}$ which is an improvement up to a power of 2 larger. Secondly, we present a technique to bound the diversity of the population that does not require a bound on its bandwidth. Apart from allowing a stronger result, we believe this is a major improvement towards the reusability of the techniques in future systematic analyses of GAs. Finally, we consider the more natural **SGA** using selection with replacement rather than without replacement although the results hold for both algorithmic versions. Experiments are presented to explore the limits of the new and previous mathematical techniques.

Keywords: Simple Genetic Algorithm, Crossover, Runtime Analysis

1. Introduction

For many years it has been a challenge to analyze the time complexity of Genetic Algorithms (GAs) using stochastic selection together with crossover and mutation. We have recently presented a first step towards a systematic analysis of GAs through a runtime analysis of the *Simple Genetic Algorithm* (**SGA**) for OneMax (Oliveto and Witt, 2012). The main result was the proof that the **SGA** has exponential runtime with overwhelming probability for population sizes up to $\mu \leq n^{1/8-\varepsilon}$ for some arbitrary small constant ε and problem size n.

The main novelties of the work were two. On one hand, we provided a rigorous proof that the **SGA** cannot optimize OneMax in polynomial time (1). The inefficient hillclimbing performance of the **SGA** due to the loss of selection pressure was well known in the

☆An extended abstract of this paper without full proofs appeared in the Proceedings of the fifteenth annual conference on Genetic and evolutionary computation (GECCO '13) (Oliveto and Witt, 2013)

Email addresses: P.Oliveto@sheffield.ac.uk (Pietro S. Oliveto), cawi@dtu.dk (Carsten Witt)

[1]Supported in part by EPSRC grant EP/H028900/1.

Evolutionary Computation (EC) community since the algorithm has been well studied in the literature. In fact, Goldberg (1989) reports experimental results in his seminal book showing the loss of selection pressure of the algorithm and suggesting fitness scaling mechanisms to solve the problem. Nevertheless a rigorous proof was yet not available. On the other hand, the major driving force was to obtain a first basis of mathematical techniques towards systematic runtime analyses of GAs using at the same time mutation, crossover and stochastic selection (2). Undoubtely significant progress has been achieved in recent years in the runtime analysis of EAs (Auger and Doerr, 2011; Jansen, 2013). Nowadays, the performance of simple EAs can be analyzed on well-known combinatorial optimization problems (Neumann and Witt, 2010). Furthermore, major advances have been achieved in the analysis of population-based EAs with stochastic selection through techniques such as the *simplified negative-drift theorem* (Oliveto and Witt, 2011), the *negative drift in populations* theorem (Lehre, 2010) and the *fitness levels for non-elitist populations* technique (Lehre, 2011). However, these techniques cannot be directly applied to the analyses of more realistic GAs incorporating a crossover operator. Several results were indeed available proving that crossover is useful (Jansen and Wegener, 2005; Watson and Jansen, 2007; Oliveto et al., 2008; Doerr et al., 2010; Kötzing et al., 2011; Neumann et al., 2011; Sudholt, 2012; Doerr et al., 2013), but they rely heavily on elitist selection operators. Moreover, mostly only upper bounds on the running time of crossover-based algorithms were available.

In this paper we present an improved runtime analysis of the **SGA** as a first step towards overcoming the limitations of our previous analysis. The first limitation was the bound on the population size $\mu = O(n^{1/8-\varepsilon})$ for the results to hold. The analysis presented here allows population sizes up to a power of 2 larger. Another significant limitation was the necessity for a bound on the so-called *bandwidth* of the population from which a measure on the diversity of the population was derived. The bandwidth was defined as $h - \ell$ where h is the best OneMax value in a population and ℓ the worst OneMax value, while diversity s was defined as the number of *non-converged* bit positions, that is both bit values are taken by individuals of the population. The whole analysis depended on the fact that if the diversity of the population is sufficiently low, then the behavior of fitness proportional selection is very close to that of uniform selection. The crucial observation to derive a bound on the diversity s was that $h - \ell \leq s$, i. e., the bandwidth cannot be larger than the number of non-converged bits. However, it is a non-trivial task to achieve a bound on the bandwidth for not too small population sizes μ. Furthermore, the bandwidth is heavily dependent on the problem at hand which seriously limits the generality and reusability of the technique. In this paper we present a new technique to bound the diversity of the population that does not require a bound on its bandwidth. Apart from allowing a stronger runtime result, we believe that this constitutes a major improvement towards the reusability of the presented technique in future analyses of the **SGA**.

Roughly speaking, in our previous work we measured the number of one-bits that individuals of the population have at a given position i in the bitstring at time t with a random variable X_t^i and showed that, if the population size is not too large, X_t^i has a very similar behavior to that of a martingale (i. e., the expected value of the random variable remains the same from one step to the next; see Williams (1991) for an introduction to martingales).

This enabled us to define a potential function $Y_t^i := (X_t^i - \mu/2)^2$ exhibiting a positive drift (i. e., the bits converge). However, for larger population sizes the positive drift does not necessarily hold since the X_t^i-process starts to resemble a submartingale (i. e., the expected value of the random variable can increase from one step to the next) and the Y_t^i-process might drift towards 0 even if the X_t^i-process increases. The proof strategy presented herein defines a different potential function Y_t^i such that a positive drift can be proved even if the underlying X_t^i-process closely resembles a submartingale. This allows the proof of exponential runtime up to population sizes $\mu \leq n^{1/4-\varepsilon}$. From the analysis an intuition can be derived that for larger populations (i. e., $\mu = \Omega(\sqrt{n})$) the bit positions no longer converge sufficiently disabling the effectiveness of the mathematical techniques presented herein and in our previous work (Oliveto and Witt, 2012). To this end, in aid of future research, we present some experiments showing how the diversity increases rapidly when the population size reaches values around $\mu = c\sqrt{n}$ for various constants $c > 0$.

The final improvement compared to our previous work (Oliveto and Witt, 2012) is a slight variation in the algorithm. We change the selection operator to select individuals from the population *with replacement* rather than *without replacement*. We feel that the chosen selection operator is the more natural variant, hence redefined the algorithm. In any case, since the probability of choosing an individual for selection twice is $O(1/\mu)$, this does not really affect the analysis. The results in the paper would also hold for the variant in Oliveto and Witt (2012), as could that variant also be used in this paper.

The rest of the paper is structured as follows. In Section 2 we discuss previous related work, define the **SGA** precisely and outline the new proof strategy in greater detail. In Section 3.1 we discuss the submartingale property of the random variable X_t^i and present a lower bound on the drift of the new potential function Y_t^i. Using the drift we derive an upper bound on the time for many bits to "almost converge" (i. e., to achieve low diversity) in Section 3.2. Finally in Section 3.3, we can apply the machinery from Oliveto and Witt (2012) to prove exponential time for the **SGA** with population sizes up to $\mu \leq n^{1/4-\varepsilon}$. In the Conclusions we present the experiments focusing on understanding at what population size the diversity starts increasing rapidly together with final remarks.

2. Algorithm and Proof Strategy

Our results are a continuation of previous work. Happ et al. (2008) performed the first runtime analysis of fitness proportional selection (f.p.s.) by considering only one individual and bitwise mutation. This work was extended by Neumann et al. (2009) to consider arbitrary population sizes again on a mutation-only EA. In particular, it was proved that the runtime of an EA using f.p.s. and bitwise mutation for OneMax is exponential with overwhelming probability (w. o. p.) whatever the polynomial population size. Also if the population is not too large (i. e., logarithmic in the problem size), then the algorithm cannot optimize any function with unique optimum in polynomial time w. o. p. Finally, in Oliveto and Witt (2012) we presented the first analysis of the complete **SGA** for OneMax using selection without replacement and proving exponential runtime for population sizes up to

$\mu \leq n^{1/8-\varepsilon}$ for some arbitrary small constant ε and problem size n. The well-known **SGA** with the more natural selection with replacement is displayed in Figure 1.

Algorithm (SGA) (Goldberg, 1989)

1. Create a parent population P consisting of μ individuals chosen uniformly at random.
2. $C := \emptyset$.
3. While $|C| < \mu$ do
   - **Fitness proportional selection:** Select two individuals x' and x'' from P according to fitness-proportional selection with replacement.
   - **Uniform crossover:** Create an offspring x by setting each bit $x(i) = x'(i)$ with probability 1/2 and $x(i) = x''(i)$ otherwise, for $1 \leq i \leq n$.
   - **Standard Bit Mutation:** Flip each bit $x(i)$ of x with probability $1/n$, for $1 \leq i \leq n$.
   - $C := C \cup \{x\}$.
4. Set $P := C$ and go to 2.

Figure 1: The Simple GA

The algorithm is initialised with a parent population P consisting of μ individuals chosen uniformly at random. In each generation a new population of size μ is created. Each individual for the new population is created by following three steps. First two parents x' and x'' are chosen from P by applying fitness-proportional selection (f.p.s.) with replacement twice. F.p.s. selects each individual $z \in P$ with probability $f(z)/\sum_{y \in P} f(y)$. Afterwards uniform crossover is applied to x' and x'' to generate an offspring x. Finally standard bit mutation is applied to x. By performing the selection, crossover and mutation steps μ times, the new parent population of size μ is obtained for the following generation.

In this paper we present an improved analysis of the **SGA** for OneMax. The OneMax function returns the number of ones in a bitstring of length n, i. e., formally defined as $\textsc{OneMax}(x) := \sum_{i=1}^{n} x_i$; often we simply write $|x|$ to denote the nuber of ones in a bitstring. The global optimum is the bitstring of only one-bits. In the following we outline the improved proof strategy.

Consider the stochastic process describing the random population vectors of the **SGA** on OneMax over time. The states of the process at time t, i. e., a concrete population, are mapped to the outcomes of the random variable X_t denoting the number of individuals with a one-bit at position 1 (the same analysis applies to all other positions). The old analysis in Oliveto and Witt (2012) started out with an idealized process, where individuals were selected uniformly. In this case, $\mathrm{E}[X_{t+1} \mid X_t] = X_t$, i. e., the process is a martingale. This implied that the process $Y_t := (X_t - \mu/2)^2$ was exhibiting positive drift $\mathrm{E}[Y_{t+1} - Y_t \mid X_t] = \mathrm{Var}[X_t]$. The variable drift theorem could be used to show that a Y-value of $(\mu/2)^2$ was obtained after an expected number $O(\mu \log \mu)$ steps, which meant that all individuals either had a one-bit or a zero-bit at position 1 then. The bit is called *converged* then and this property was used to bound the effect of crossover.

Furthermore, the old analysis exploited that the actual process, where fitness-proportional instead of uniform selection is used, is very close to the idealized process under strong assumptions, including the bound $\mu = O(n^{1/8-\varepsilon})$. Then the X_t were so close to a martingale that Y_t still was drifting towards its maximum value. However, for larger μ the following problem occurs: the actual process (if mutation is ignored) is a submartingale, i. e., $\mathrm{E}[X_{t+1} \mid X_t] \geq X_t$, i. e., there is positive drift $\mathrm{E}[X_{t+1} - X_t \mid X_t] > 0$. Unless the strong assumptions were made, the drift of the X_t-process was too strong to maintain positive drift of the Y_t-process. For instance, if $X_t = \mu/4$, then the positive drift of the X_t-process towards the "middle" $\mu/2$ could imply a negative drift of the Y_t-process.

We can overcome this difficulty to some extent by redefining $Y_t := X_t^2$. Our aim is to show that the Y_t-process has a drift that can be bounded from below by a useful value (whereas the drift of the X_t-process is not necessarily strong enough). The expected first hitting time for either 0 or at least $\mu^2(1 - O(1/n))^2$, i. e., the expected time for the bit to "almost" converge, will be bounded by $O(\mu \log^3 n)$. As long as $\mu = O(n^{1/4-\varepsilon})$, many almost converged bits will hinder the process from reaching the optimum. On the way towards this bound, we need:

1. a formal proof that the X_t-process resembles a submartingale (Lemma 3),
2. a lower bound on the drift of the Y_t-process (Lemma 4),
3. an application of the variable drift theorem (Lemma 7). Here the additional difficutly arises that the drift is not monotone in the distance to the optimum, which is required by the standard version of the variable drift theorem. Fortunately, there is a more recent generalization of the variable drift theorem in Feldmann and Kötzing (2013) that can be adapted to our case. In order to apply it, some technical conditions have to be verified, e. g., that the maximum jump size of the Y_t-process is bounded (Lemma 9).

Once the new convergence analysis is obtained, we will apply the machinery from our previous work (Oliveto and Witt, 2012) to achieve the final result. Roughly speaking, a potential function collapsing the whole population into a single value is defined and a drift away from the optimum is proven, given that there is a sufficient number of converged bits; see Section 3.3.

3. Detailed Theoretical Analysis

3.1. "Submartingale" Property of X_t and Drift of Y_t

We will analyze the X_t-process and find out that $\mathrm{E}[X_{t+1} \mid X_t] \geq X_t(1 - 2/n)$, see Lemma 3. If the term $-2/n$ (which is due to mutation) was not present, this would characterize the process as a submartingale. As $\mu = o(n)$ and $X_t \leq \mu$, the term $-2/n$ is not significant, which intuitively means that the process still resembles a submartingale. To prove the desired property, we need a helper result, namely Lemma 2. There we analyze the stochastic process describing the populations induced by the **SGA** on OneMax. To formalize this, let the vector $P_t = \{(x_1^{(t)}, \ldots, x_\mu^{(t)})\}$, where $t \geq 0$ and $x_j^{(t)} \in \{0,1\}^n$ for $j \in \{1, \ldots, \mu\}$, be the population at time t, ordered in an arbitrary but fixed way. Note that

this is a random vector, whose components are uniform at random at time 0. At later points of time, the components are not necessarily uniform due to the **SGA** favoring individuals with larger number of one-bits. However, we observe that the positions of the one-bits are still uniformly distributed, as the following lemma formalizes.

Lemma 1. *Consider some population $P_t = \{(x_1^{(t)}, \ldots, x_\mu^{(t)})\}$, where $t \geq 0$, of the* **SGA** *on* OneMax. *Let $j \in \{1, \ldots, \mu\}$ be arbitrary but fixed and let $x := x_j^{(t)}$. Let $(x_1, \ldots, x_n)$ denote the bit vector given by x. If $|x| = k$, then for any bit index $i \in \{1, \ldots, n\}$, it holds* $\text{Prob}(x_i = 1) = k/n$.

Proof. The claim is equivalent to that x is uniformly distributed on the set $V_k := \{v \in \{0,1\}^n \mid |v| = k\}$. We prove it by induction. Obviously, it holds for $t = 0$ since all individuals are drawn uniformly from $\{0,1\}^n$. The transition from time t to time $t+1$ consists of selection, crossover, and mutation. Selection chooses based on the number of one-bits, only, since the objective function value does not depend on the position of the one-bits. Hence, all individuals with k one-bits have the same selection probability. The mutation and crossover operator used are so-called unbiased ones (Lehre and Witt, 2012). Formally, consider any permutation $\pi \in S_n$, where S_n denotes the set of permutations on $\{1, \ldots, n\}$ and define the permutation of a bitstring $x = (x_1, \ldots, x_n)$ by $\pi(x) = (\pi(x_1), \ldots \pi(x_n))$. Then, with the stochastic mappings $\text{mut}\colon \{0,1\}^n \to \{0,1\}^n$ and $\text{cross}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^n$ induced by mutation and crossover, respectively, we have $\text{Prob}(\text{mut}(x) = z) = \text{Prob}(\text{mut}(\pi(x)) = \pi(z))$ as well as $\text{Prob}(\text{cross}(x,y) = z) = \text{Prob}(\text{cross}(\pi(x), \pi(y)) = \pi(z))$ for all $x, y, z \in \{0,1\}^n$.

The inductive assumption says that if $|x| = k$ for the individual selected for mutation, then $\text{Prob}(x = v^*) = 1/\binom{n}{k}$ for every $v^* \in V_k$. Let $y \in \{0,1\}^n$ be arbitrary and denote $p^* := \text{Prob}(\text{mut}(x,y))$. We get that $\text{Prob}(\text{mut}(\pi(x), \pi(y))) = p^*$ for every $\pi \in S_n$ and note that $\{\pi(y) \mid \pi \in S_n\}$ contains all individuals in $S_{|y|}$ one-bits. Since x is uniform on S_k and π is a bijection, we get that $\text{mut}(x,y)$ assigns uniform probability to all individuals with the same number of one-bits. Analogously, the statement is proved for crossover. This proves the induction step. □

We can now formulate the claim that the number of one-bits at an arbitrary fixed position (w. l. o. g., position 1) determines a lower bound on the probability of selecting an individual having a one there.

Lemma 2. *Consider the random population vector of the* **SGA** *on* OneMax *at an arbitrary point of time. Let k be the number of individuals with a one-bit at position 1. Then an application of the fitness-proportional selection operator will select such an individual with probability at least k/μ.*

Proof. Let a random population $P = (x_1, \ldots, x_\mu)$ at the considered point in time be given, ordered in an arbitrary but fixed way. Let I be the random variable denoting the number of individuals with a one-bit at position 1. Note that the theorem conditions on $I = k$; however, for the moment we drop this condition and let I be random.

All x_i, where $1 \leq i \leq \mu$, so far follow the same probability distribution. By Lemma 1, the one-bits of x_i (note that also $|x_i|$ is a random variable) are uniformly distributed among the individual. We investigate the consequences of introducing the condition that $f(x_i) \geq f(x_j)$ for some $j \neq i$, which is equivalent to $|x_i| \geq |x_j|$. This event applies to the sum of the bit values and not a particular bit. Hence, Lemma 1 still applies under the assumption $|x_i| \geq |x_j|$ (recall that we do not condition on $I = k$ yet). Denoting by p_i the probability that individual x_i has a one-bit at position 1, we have both $p_i = \mathrm{E}[|x_i|]/n$ and $p_j = \mathrm{E}[|x_j|]/n$, which implies $p_i \geq p_j$.

What happens if we additionally condition on $I = k$ for some fixed k? First of all, note that the one-bits of an individual are no longer guaranteed to be uniformly distributed. Secondly, the statement of the lemma is trivial if either $k = 0$ or $k = \mu$. In the following, we shall consider the condition that that $I = k$ only for $k \in \{1, \ldots, \mu - 1\}$. Recalling that p_i denotes the probability of having a one at position 1 in individual x_i before conditioning, we define $p'_i = (p_i \mid I = k)$ for $i \in \{1, \ldots, \mu\}$. We again consider two individuals x_i and x_j and condition on $|x_i| \geq |x_j|$ in the rest of this paragraph. The aim is to show that $p'_i \geq p'_j$. By definition of conditional probability,

$$p'_i = \frac{\mathrm{Prob}(x_i \text{ has a one at position 1 and } I = k \text{ happens})}{\Pr(I = k)}$$

and accordingly for p'_j. Let $\tilde{X} := \{x_1, \ldots, x_\mu\} \backslash \{x_i, x_j\}$ denote the $\mu - 2$ remaining individuals and R denote the number of one-bits at position 1 in $\tilde{X}$. The event $I = k$ is only possible if $R \in \{k-2, k-1, k\}$. We further condition on the value of R. If $R = k$ then immediately $p'_i = p'_j = 0$; similarly if $R = k-2$ then $p'_i = p'_j = 1$. Finally, if $R = k-1$, then the condition $(I = k) \wedge (R = k - 1)$ implies that exactly one of the two individuals has a one-bit at position 1. By Lemma 1, before conditioning on concrete values for I and R, the probability of x_i getting a one and x_j a zero is

$$p_{i \wedge \bar{j}} := \frac{|x_i|}{\mu} \left(1 - \frac{|x_j|}{\mu}\right)$$

and the reverse case has probability

$$p_{j \wedge \bar{i}} := \frac{|x_j|}{\mu} \left(1 - \frac{|x_i|}{\mu}\right).$$

We claim that

$$\begin{aligned}
(p'_i \mid R = k-1) &= (p_i \mid (I = k) \wedge (R = k-1)) \\
&= \frac{p_{i \wedge \bar{j}} \cdot \mathrm{Prob}(\tilde{X} \text{ has } k-1 \text{ ones at pos. } 1 \mid x_i \text{ has a one and } x_j \text{ a zero there})}{\Pr((I = k) \wedge (R = k-1))} \\
&= \frac{p_{i \wedge \bar{j}} \cdot \mathrm{Prob}(\tilde{X} \text{ has } k-1 \text{ ones at pos. } 1 \mid x_j \text{ has a one and } x_i \text{ a zero there})}{\Pr((I = k) \wedge (R = k-1))}. \qquad (*)
\end{aligned}$$

Note that, before conditioning, all individuals in P are identically and independently distributed. We inspect the event from the second term in the numerator of Equation $(*)$, i. e., the event that $\tilde{X}$ has $k-1$ ones as position 1 conditioned on that both $|x_i| \geq |x_j|$ and that x_i has a one and x_j a zero there. Now, by conditioning on certain properties for only x_i and x_j, more precisely $|x_i| \geq |x_j|$ and that x_i has a one-bit at position 1 and x_j has not, we do not introduce dependencies among the creation of the individuals in $\tilde{X}$. Hence, the distribution of the one-bits in every individual from $\tilde{X}$ does not depend on whether x_i or x_j has the one-bit such that Equation $(*)$ follows.

We obtain an expression analogue to Equation $(*)$ for $(p'_j \mid R = k-1)$. Therefore $p'_i/p'_j = p_{i\wedge\bar{j}}/p_{j\wedge\bar{i}}$ on $R = k$ for every $k \in \{0, \ldots, \mu\}$. Now, since $|x_i| \geq |x_j|$, both

$$\frac{|x_i|}{\mu} \geq \frac{|x_j|}{\mu} \quad \text{and} \quad \left(1 - \frac{|x_j|}{\mu}\right) \geq \left(1 - \frac{|x_i|}{\mu}\right),$$

and $p'_i \geq p'_j$ follows.

Next we consider the selection probabilities. Let s'_i be the probability that fitness-proportional selection selects x_i, already conditioning on $I = k$. By definition of the selection operator,

$$s'_i = \mathrm{E}\left[\frac{f(x_i)}{\sum_{j=1}^{\mu} f(x_j)} \mid x_1, \ldots, x_\mu\right].$$

Hence, if $|x_i| \geq |x_j|$, we get $s'_i \geq s'_j$. The condition $I = k$ does not introduce any complications at this point.

Also if we extend the conditioning on the order of fitness values to several individuals, e. g., by assuming $f(x_{i_1}) \geq f(x_{i_2}) \geq \cdots \geq f(x_{i_r})$ for indices $i_1, \ldots, i_r$, where $r \leq \mu$, the same reasoning as before applies. Consequently, if we consider the order statistics of the fitness values by sorting the individuals as $x_{i_1}, \ldots, x_{i_\mu}$ according to decreasing fitness such that $f(x_{i_1}) \geq \cdots \geq f(x_{i_\mu})$, we have that $s'_{i_1} \geq \cdots \geq s'_{i_\mu}$ and $p'_{i_1} \geq \cdots \geq p'_{i_\mu}$.

Let p^* be the probability of selecting an individual with a one-bit at position 1 in a single selection trial, which is the probability we have to bound in the lemma. By definition, $p^* = s'_{i_1}p'_{i_1} + \cdots + s'_{i_\mu}p_{i_\mu}$. Now, $s'_{i_1} + \cdots + s'_{i_\mu} = 1$ since the s'_{i_j} form a probability distribution. Hence, Chebyshev's sum inequality yields $p^* \geq (p'_{i_1} + \cdots + p'_{i_\mu})/\mu$. Now, by using indicator random variables for the events associated with the p'_{i_j} and linearity of expectation, we get the identity $p'_{i_1} + \cdots + p'_{i_\mu} = \mathrm{E}[I]$. Since we still condition on $I = k$, we have $\mathrm{E}[I] = k$ and altogether $p^* \geq k/\mu$. □

We now prove the "almost-submartingale" property. For technical reasons, we will also have to analyze a modified **SGA**, where selection is uniform, i. e., does not take fitness into account.

Lemma 3. *The random variable X_{t+1} is the sum of μ independent Bernoulli trials with common success probability p, where $p \geq (X_t/\mu)(1-2/n) + 1/n$. Moreover, $\mathrm{E}[X_{t+1} \mid X_t] \geq X_t(1-2/n) + \mu/n$.*

If the **SGA** *uses uniform instead of fitness-proportional selection, then $p = (X_t/\mu)(1-2/n) + 1/n$.*

Proof. We consider the stochastic process describing the random populations of the **SGA** on OneMax over time and inspect the creation of a single individual when going from time t to the next generation. For the individual to have a one-bit (at position 1), either the result of crossover must be a one-bit and mutation must not flip it, or the result of crossover must be a zero-bit and mutation must flip it. The result of crossover is a one-bit for sure if both parents have a one-bit, and it is a one-bit with probability 1/2 if the parents differ at the position (which happens if the first parent has a one and the second a zero or the other way round). Let s be the probability of fitness-proportional choosing an individual with a one-bit. We get

$$\begin{aligned} p &= s(1-s)\frac{1}{2}\left(1-\frac{1}{n}\right) + (1-s)s\frac{1}{2}\left(1-\frac{1}{n}\right) + s^2\left(1-\frac{1}{n}\right) \\ &\quad + s(1-s)\frac{1}{2}\frac{1}{n} + (1-s)s\frac{1}{2}\frac{1}{n} + (1-s)^2\frac{1}{n} \\ &= s(1-s) + s^2 - \frac{1}{n}(s^2-(1-s)^2) = s - \frac{2s-1}{n} = s\left(1-\frac{2}{n}\right) + \frac{1}{n}. \end{aligned}$$

Note that the outcomes of the random variables X_t are unambigously obtained from the state of the stochastic process mentioned above. The bound on p follows since $s \geq X_t/\mu$ according to Lemma 2. Obviously, $s = X_t/\mu$ in case of uniform selection, so that the bound becomes an equality then.

The bound on $\mathrm{E}[X_{t+1} \mid X_t]$ follows immediately from the bound on p since the number of trials equals μ. □

Now we consider the new process $Y_t = X_t^2$ with the aim to prove that it drifts. Later, this drift will allow us to bound the time for bits to converge.

Lemma 4. *For $t \geq 0$ it holds*

$$\begin{aligned} \mathrm{E}[X_{t+1}^2 - X_t^2 \mid X_t] &\geq \mathrm{E}[(X_{t+1} - (X_t - 2X_t/n))^2 \mid X_t] - 4X_t^2/n \\ &\geq \mathrm{E}[(X_{t+1} - \mathrm{E}[X_{t+1}])^2 \cdot \mathbb{1}\{X_{t+1} \geq \mathrm{E}[X_{t+1}]\} \mid X_t] - 4X_t^2/n. \end{aligned}$$

Proof. We first study the second inequality in the statement of the lemma. That is,

$$\begin{aligned} &\mathrm{E}[(X_{t+1} - (X_t - 2X_t/n))^2 \mid X_t] - 4X_t^2/n \\ &\qquad \geq \mathrm{E}[(X_{t+1} - \mathrm{E}[X_{t+1}])^2 \cdot \mathbb{1}\{X_{t+1} \geq \mathrm{E}[X_{t+1}]\} \mid X_t] - 4X_t^2/n. \end{aligned}$$

Since the first term on the left-hand side is squared, hence non-negative, the inequality holds iff

$$\mathrm{E}[(X_{t+1} - (X_t - 2X_t/n))^2 \mid X_t] \geq \mathrm{E}[(X_{t+1} - \mathrm{E}[X_{t+1}])^2 \cdot \mathbb{1}\{X_{t+1} \geq \mathrm{E}[X_{t+1}]\} \mid X_t] \qquad (1)$$

If the indicator function on the right hand side is zero, then Inequality (1) holds trivially. So we study the inequality when the indicator function is non-zero, i.e., $X_{t+1} \geq \mathrm{E}[X_{t+1} \mid X_t]$. The aim is to show that the term to-be-squared on the left-hand side is non-negative and

greater than the term to-be-squared on the right-hand side. Then the inequality with the squared terms will also follow.

Since by Lemma 3, $\mathrm{E}[X_{t+1} \mid X_t] \geq X_t(1-2/n) + \mu/n \geq X_t(1-2/n)$, it follows that $X_{t+1} \geq \mathrm{E}[X_{t+1} \mid X_t] \geq X_t - 2X_t/n$. Hence, the expression to be squared on the left side of Inequality (1) is non-negative if the indicator function is non-zero, i.e., $X_{t+1} - (X_t - 2X_t/n) \geq 0$.

By using Lemma 3 again (i.e., $\mathrm{E}[X_{t+1} \mid X_t] \geq X_t - 2X_t/n$), we obtain $X_{t+1} - (X_t - 2X_t/n) \geq X_{t+1} - \mathrm{E}[X_{t+1} \mid X_t]$ and the desired inequality follows since the term on the left-hand side to be squared is non-negative.

The first inequality follows from this by elementary manipulations. More precisely,

$$\begin{aligned}
&\mathrm{E}[(X_{t+1} - (X_t - 2X_t/n))^2 \mid X_t] \\
&\qquad = \mathrm{E}[X_{t+1}^2 \mid X_t] - \mathrm{E}[2X_{t+1}(X_t - 2X_t/n) - (X_t - 2X_t/n)^2 \mid X_t] \\
&\qquad \leq \mathrm{E}[X_{t+1}^2 \mid X_t] - \mathrm{E}[2(X_t - 2X_t/n)^2 - (X_t - 2X_t/n)^2 \mid X_t] \\
&\qquad \leq \mathrm{E}[X_{t+1}^2 \mid X_t] - \mathrm{E}[X_t^2 \mid X_t] + 4X_t^2/n,
\end{aligned}$$

where the first estimation used Lemma 3 again. □

The aim is now to show in Lemma 6 that X_{t+1} with at least constant probability exceeds its expectation by a considerable amount, which, together with Lemma 4, will allow us to bound the drift of the Y-process (Lemma 7). This is contrary to the typical applications of Chernoff-Hoeffding bounds, where the aim is to get an upper bound on the probability of a random variable exceeding its expectation considerably.

To show the lower bound on the probability of a deviation, we make use of the following well-known result from probability theory (Feller, 1971), which has been rarely used in the theory of randomized search heuristics before. We apply it afterwards in Lemma 6.

Lemma 5 (Berry-Esseen inequality)**.** *Let $X_1, \ldots, X_\mu$ be independent, identically distributed random variables with $\mathrm{E}[X_i] = 0$, $\mathrm{E}[X_i^2] = \sigma^2 > 0$, and $\mathrm{E}[|X_i|^3] = \rho < \infty$ for $1 \leq i \leq \mu$. Then there is a constant $C > 0$ such that the cumulative distribution function (cdf.) F_Y of $Y := (X_1 + \cdots + X_\mu)/(\sigma\sqrt{\mu})$ satisfies*

$$|F_Y(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{\mu}}$$

for all $x \in \mathbb{R}$ (where $\Phi(x)$ denotes the cdf. of the standard normal distribution).

Lemma 6. *Let X be the sum of μ independent Bernoulli trials with success probability r/μ each, for some $r \in [0, \mu]$. Let $\ell := \min\{r, \mu - r\}$. Then*

1. $\mathrm{Prob}[X \geq \mathrm{E}[X] + \sqrt{\ell/2}] = \Omega(1)$.
2. $\mathrm{E}[(X - \mathrm{E}[X]) \cdot \mathbb{1}\{X \geq \mathrm{E}[X]\}] \leq \sqrt{\ell}$.

Proof. Within this proof, we let X_i denote the $\{0,1\}$-random variable corresponding to the i-th trial, $1 \le i \le \mu$. The aim is to apply Lemma 5. We introduce $X_i' := X_i - r/\mu$ and note that

$$\begin{aligned}
\mathrm{E}[X_i'] &= \frac{r}{\mu} - \frac{r}{\mu} = 0 \\
\sigma^2 := \mathrm{E}[X_i'^2] &= \frac{r}{\mu}\left(1-\frac{r}{\mu}\right)^2 + \left(1-\frac{r}{\mu}\right)\left(-\frac{r}{\mu}\right)^2 = \frac{r}{\mu}\left(1-\frac{r}{\mu}\right) \\
\rho := \mathrm{E}[|X_i'|^3] &= \frac{r}{\mu}\left(1-\frac{r}{\mu}\right)^3 + \left(1-\frac{r}{\mu}\right)\left(\frac{r}{\mu}\right)^3 = \frac{r}{\mu}\left(1-\frac{r}{\mu}\right)\left(\left(1-\frac{r}{\mu}\right)^2 + \left(\frac{r}{\mu}\right)^2\right) \\
&\le \frac{r}{\mu}\left(1-\frac{r}{\mu}\right).
\end{aligned}$$

Consider the random variable Y as defined in Lemma 5 and note that $\sigma\sqrt{\mu} = \sqrt{r(\mu-r)/\mu} \ge \sqrt{\min\{r,\mu-r\}/2}$. Hence, $X \ge r + \sqrt{\ell/2} = \mathrm{E}[X] + \sqrt{\ell/2}$ is implied by $Y \ge \mathrm{E}[Y] + 1$. The lemma yields

$$\mathrm{Prob}[Y \le \mathrm{E}[Y]+1] \le \Phi(1) + \frac{C\rho}{\sigma^3\sqrt{\mu}} \le \Phi(1) + \frac{C\sqrt{\mu}}{\sqrt{r(\mu-r)}} \le \Phi(1) + \frac{\sqrt{2}\cdot C}{\sqrt{\ell}}.$$

If ℓ is greater than a sufficiently large constant, then the last bound is less than 1 and the theorem is proved in this case. Otherwise, we have $\ell = O(1)$ and distinguish between two subcases. If $r \ge \mu/2$, we obtain $\mathrm{Prob}[X = \mu] \ge (1-\ell/\mu)^\mu = \Omega(1)$. Otherwise, we note that $\mathrm{Prob}[X = s] \ge \binom{\mu}{s}(\ell/\mu)^s(1-\ell/\mu)^s = \Omega(1)$ for any constant $s \ge 0$. Hence, $\mathrm{Prob}[X \ge r + \sqrt{\ell/2}] = \Omega(1)$ in any case. This proves the first statement of the lemma.

For the second statement, we note that

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[(X-\mathrm{E}[X])^2 \cdot \mathbb{1}\{X \ge \mathrm{E}[X]\}] + \mathrm{E}[(X-\mathrm{E}[X])^2 \cdot \mathbb{1}\{X < \mathrm{E}[X]\}] \\
&\ge (\mathrm{E}[(X-\mathrm{E}[X]) \cdot \mathbb{1}\{X \ge \mathrm{E}[X]\}])^2
\end{aligned}$$

by Jensen's inequality. Hence,

$$\mathrm{E}[(X-\mathrm{E}[X]) \cdot \mathbb{1}\{X \ge \mathrm{E}[X]\} \le \sqrt{\mathrm{Var}[X]} = \sqrt{\frac{r}{\mu}(\mu-r)} \le \sqrt{\min\{r,\mu-r\}}$$

as claimed. □

We can now bound the drift of the Y-process if μ is not too large.

Lemma 7. *For $t \ge 0$, $\mathrm{E}[Y_{t+1} - Y_t \mid X_t] = \Omega(\min\{X_t, \mu - X_t\}) - 4X_t^2/n$. Moreover, for $\mu = o(\sqrt{n})$ and $X_t \le \mu - 1$, $\mathrm{E}[Y_{t+1} - Y_t \mid X_t] = \Omega(\min\{X_t, \mu - X_t\})$.*

Proof. Recall that $Y_t = X_t^2$ for all $t \ge 0$. According to Lemma 3, X_{t+1} stochastically dominates a random variable following the binomial distribution with parameters μ and

$(X_t/\mu)(1-2/n)$. Moreover, according to Lemma 4, it is sufficient to establish a deviation of X_{t+1} above its mean in order to bound the drift of the Y-process. Applying Lemma 6 with $r = X_t(1-2/n)$, the event

$$X_{t+1} \geq \mathrm{E}[X_{t+1}] + \sqrt{\frac{\min\{X_t(1-2/n), \mu - X_t(1-2/n)\}}{2}}$$

occurs with probability at least c for some constant $c > 0$ and sufficiently large n. Moreover, $1 - 2/n \geq 1/2$ for sufficiently large n. Hence,

$$\mathrm{E}[(X_{t+1} - \mathrm{E}[X_{t+1}])^2 \cdot \mathbb{1}\{X_{t+1} \geq \mathrm{E}[X_{t+1}]\} \mid X_t) \;\geq\; (c/4)\min\{X_t, \mu - X_t\}$$

for sufficiently large n, such that the statement of Lemma 4 completes the proof of the first statement in the lemma here.

A sufficient condition for the second statement to follow from the first one is given by $4X_t^2/n \leq (c/8)\min\{X_t, \mu - X_t\}$. Since $\mu = o(\sqrt{n})$ and $X_t \leq \mu - 1$, $4X_t^2/n \leq (c/8)X_t$ is satisfied for sufficiently large n. Since $X_t \leq \mu - 1$, the condition $4X_t^2/n \leq (c/8)(\mu - X_t)$ follows from $4X_t^2/n \leq c/8$, which again follows for sufficiently large n since $\mu = o(\sqrt{n})$. □

So far, we have established a positive drift of the Y-process as long as $X_t \leq \mu - 1$. This drift is dependent on X_t. Unlike the analysis in Oliveto and Witt (2012), the drift is not monotone in X_t. Therefore, the standard assumption of the variable drift theorem (see Rowe and Sudholt, 2012 for the most recent version) does not hold. We use the following variant (correcting a minor mistake in a formulation by Feldmann and Kötzing, 2013) instead. Its proof is given in Appendix B.

Theorem 1 (extending Feldmann and Kötzing, 2013)**.** *Let $(Z_t)_{t\geq 0}$, be a stochastic process adapted to some filtration $\mathcal{F}_t$ over a state space $S \subseteq \{0\} \cup [z_{\min}, z_{\max}]$, where $z_{\min} > 0$. Suppose there exist two functions function $h, d\colon [z_{\min}, z_{\max}] \to \mathbb{R}^+$, where $1/h$ is integrable, and some $c \geq 1$ not depending on Z_t such that for all $t \geq 0$*

(1) $\mathrm{E}[Z_t - Z_{t+1} \mid \mathcal{F}_t; Z_t \geq z_{\min}] \geq h(Z_t)$,
(2) $\frac{\mathrm{E}[(Z_t - Z_{t+1})\cdot \mathbb{1}\{Z_{t+1} < Z_t\} \mid \mathcal{F}_t; Z_t \geq z_{\min}]}{\mathrm{E}[(Z_{t+1} - Z_t)\cdot \mathbb{1}\{Z_{t+1} > Z_t\} \mid \mathcal{F}_t; Z_t \geq z_{\min}]} \geq 2c^2$,
(3) $|Z_t - Z_{t+1}| \leq d(Z_t)$.
(4) *for all $x, y \geq z_{\min}$ with $|x - y| \leq d(x)$ it holds $h(\min\{x, y\}) \leq c \cdot h(\max\{x, y\})$.*

Then it holds for the first hitting time $T := \min\{t \mid Z_t = 0\}$ that

$$\mathrm{E}[T \mid Z_0] \leq 2c\left(\frac{z_{\min}}{h(z_{\min})} + \int_{z_{\min}}^{Z_0} \frac{1}{h(x)}\,\mathrm{d}x\right).$$

We will apply the previous theorem w. r. t. the process $Z_t := \mu^2 - Y_t = \mu^2 - X_t^2$, where the aim is to minimize the Z-value. According to Lemma 7, we get

$$\mathrm{E}[Z_t - Z_{t+1} \mid X_t] = \Omega(\min\{\sqrt{\mu^2 - Z_t}, \mu - \sqrt{\mu^2 - Z_t}\})$$

if $1 \leq X_t \leq \mu - 1$.

Loosely speaking, the drift is a parabolic function with a maximum at $Z_t = 3\mu^2/4$. In particular, it is monotone increasing in Z_t only if $Z_t < 3\mu^2/4$. In order to apply Theorem 1, we need a bound on the maximum change of the process and the ratio of positive and negative drift. Unfortunately, we only have a lower bound on p according to Lemma 3 and cannot control the size of jumps from X_t towards the optimum (i. e., state μ for X_t, which is the same as state 0 for Z_t).

To overcome this problem, we consider the above-mentioned modified **SGA** using uniform instead of fitness-proportional selection and denote by $\tilde{X}_t$ its current number of one-bits at a fixed position at time t. Lemma 3 tells us the upper bound $p \leq \tilde{X}_t/\mu + 1/n$ for this process. Intuitively, the $\tilde{X}_t$-process should need longer time to hit state μ than the actual one. However, we are not really interested in the first hitting time T_μ of μ but rather in the first hitting time $T_{0 \vee \mu}$ of either 0 or μ, which are the states corresponding to a converged bit. Obviously, T_μ stochastically dominates $T_{0 \vee \mu}$; however, T_μ might be a very bad or even useless estimate for $T_{0 \vee \mu}$ if the underlying process is unlikely to leave state 0. We can allow ourselves to modify the transition probabilities from state 0 in both the original X_t-process and the $\tilde{X}_t$-process without changing the corresponding $T_{0 \vee \mu}$ and then use T_μ for the modified process as a bound on $T_{0 \vee \mu}$. Moreover, we can make state μ absorbing without changing any of the first hitting times.

From now on, both processes are permanently modified at state 0 by letting $X_{t+1} = \mathrm{Bin}(\mu, (1 - 2/n)/\mu + 1/n)$ if $X_t = 0$ and accordingly for $\tilde{X}_t$. This means that the transition probabilities from state 0 are set to those from state 1. Moreover, both processes are changed to transit from state μ to state μ with probability 1. The resulting processes are called X'_t and $\tilde{X}'_t$ hereinafter.

The first hitting time of μ for the X'_t-process is an upper bound (in the sense of stochastic dominance) on the time for a bit to converge, i. e., for the completely unmodified process to reach state 0 or μ. We claim that the first hitting time of μ for the $\tilde{X}'_t$-process is an upper bound on the corresponding first hitting time for the X'_t-process. This is made rigorous by the following lemma, which uses $\succ$ to denote stochastic dominance of random variables and T' and $\tilde{T}'$ to denote the first hitting time of μ for the X'-process and the $\tilde{X}'$-process, respectively.

Lemma 8. *For $t \geq 0$, $X'_t \succ \tilde{X}'_t$. Moreover, $\tilde{T}' \succ T'$.*

Proof. The first claim is proved inductively. It obviously holds at time 0 since both processes are initialized in the same way. By Lemma 3, if the state at time t is less than μ,

$$X'_{t+1} \succ \mathrm{Bin}\left(\mu, \frac{\max\{1, X'_t\}}{\mu}\left(1 - \frac{2}{n}\right) + \frac{1}{n}\right),$$
$$\tilde{X}'_{t+1} = \mathrm{Bin}\left(\mu, \frac{\max\{1, \tilde{X}'_t\}}{\mu}\left(1 - \frac{2}{n}\right) + \frac{1}{n}\right),$$

where $\mathrm{Bin}(a, b)$ denotes a random variable following the binomial distribution with parameters a and b; moreover, if the state at time t equals μ, then the distribution at time $t + 1$

is $\mathrm{Bin}(\mu, 1)$ for both processes. Now, since the success probability in the first binomial distribution is monotone increasing w. r. t. X'_t, we obtain from the induction hypothesis that $X'_{t+1} \succ \mathrm{Bin}\Big(\mu, \frac{\max\{1,\tilde{X}'_t\}}{\mu}\left(1-\frac{2}{n}\right)+\frac{1}{n}\Big) = \tilde{X}'_{t+1}$, which proves the induction step.

To prove the second claim, we recall that both processes stop when state μ is reached. Hence, $T' > t$ is equivalent to $X'_t < \mu$ and accordingly for the $\tilde{X}'$-process. Hence, $\mathrm{Prob}[T' > t] = 1 - \mathrm{Prob}[X'_t \geq \mu]$ and accordingly $\mathrm{Prob}[\tilde{T}' > t] = 1 - \mathrm{Prob}[\tilde{X}'_t \geq \mu]$. Using the first claim, the second one now follows. □

From now on, we abuse notation by writing X_t but in fact meaning the doubly-modified $\tilde{X}'$-process; also the Z-process is defined based on this $\tilde{X}'$-process. The following lemma is used to establish some of the prerequisites of Theorem 1.

Lemma 9. *Let $\mu = o(\sqrt{n})$. Then $\mathrm{Prob}[|X_{t+1} - X_t| \geq (X_t)^{2/3} + 1/2] \leq 2e^{-X_t^{1/3}/4}$. Moreover, $\frac{\mathrm{E}[(Z_t - Z_{t+1})\cdot \mathbb{1}\{Z_{t+1} < Z_t\} | X_t]}{\mathrm{E}[(Z_{t+1} - Z_t)\cdot \mathbb{1}\{Z_{t+1} > Z_t\} | X_t]} = 1 + \Omega(1)$.*

Proof. The first statement follows from standard Chernoff bounds since $\mathrm{E}[X_{t+1} \mid X_t] \geq X_t(1-2/n) \geq X_t - 1/2$ as well as $\mathrm{E}[X_{t+1} \mid X_t] \leq X_t + \mu/n \leq X_t + 1/2$ for n large enough (since $\mu = o(\sqrt{n})$).

For the second statement, we first assume $X_t \leq \mu/2$ and observe that

$$\begin{aligned}\mathrm{E}[(Z_t - Z_{t+1}) \cdot \mathbb{1}\{Z_{t+1} < Z_t\} \mid X_t] - \mathrm{E}[(Z_{t+1} - Z_t) \cdot \mathbb{1}\{Z_{t+1} > Z_t\} \mid X_t] &= \mathrm{E}[Z_t - Z_{t+1} \mid X_t] \\ &= \Omega(X_t)\end{aligned}$$

according to Lemma 7. Moreover, from the second statement of Lemma 6, we get $\mathrm{E}[(Z_{t+1} - Z_t) \cdot \mathbb{1}\{Z_{t+1} > Z_t\} \mid X_t] = O(X_t)$. Combining the two observations, we obtain the claim. The case $X_t \geq \mu/2$ is proved analogously with X_t replaced by $\mu - X_t$. □

Taking everything together, we can bound the time for a bit to converge.

Lemma 10. *Consider the* **SGA** *on* OneMax *and let $\mu = o(\sqrt{n})$. Then the expected time for an arbitrary bit position to reach either 0 or at least μ one-bits is $O(\mu \log^3 n)$.*

Proof. The aim is to apply drift analysis, in particular Theorem 1, on the Z_t-process defined above and to bound its first hitting time T of state 0, which, as argued in the context of Lemma 8, is a bound on the time for a bit to converge. We start with a simple case. If $\mu \leq 1458 \ln^3 n := d^*$ then $\mu^2 = O(\mu \log^3 n)$. Since $E[Z_t - Z_{t+1} \mid Z_t > 0] = \Omega(1)$ by Lemma 7, and $Z_t \leq \mu^2$, we get the bound $O(\mu^2)$ already by classical additive drift.

From now on, we assume $\mu > d^*$. One prerequisite of Theorem 1 is given by a bound on the maximum jump size, which will be stated in terms of the X_t-process. Let

$$d(x) := \begin{cases} x^{2/3} + \frac{1}{2} & \text{if } x \geq d^*/2, \\ \ell^* := 81 \ln^2 n + \frac{1}{2} & \text{otherwise.} \end{cases}$$

Note that $\ell^* = (d^*/2)^{2/3} + 1/2$. If $X_t \geq d^*/2$, we get from the first statement of Lemma 9 that

$$\mathrm{Prob}[|X_{t+1} - X_t| \geq d(X_t)] \leq 2e^{-(9/4)\ln n} = 2n^{-9/4}.$$

If $X_t < d^*/2$, then we still get $\text{Prob}[X_{t+1} \geq X_t + d(X_t)] \leq 2n^{-9/4}$ by using $d^*/2$ as an upper bound on X_t. To analyze the deviation in the other direction, we use the symmetry of the binomial distribution, more precisely, $\text{Prob}[X_{t+1} \leq X_t - d(X_t)] = \text{Prob}[X_{t+1} \geq (\mu - X_t) + d(X_t)]$. Now, since $\mu - X_t \geq d^*/2$, the bound $2n^{-9/4}$ follows again as before. Throughout this proof, we assume $|X_{t+1} - X_t| \leq d(X_t)$ for a period of $\Theta(\mu \log^3 n)$ steps. By a union bound, the failure probability is $o(n) \cdot n^{-9/4} = O(n^{-5/4})$. Moreover, as the maximum change of the Z_t-process is $\mu^2 = o(n)$, the jumps of larger size can contribute only $o(n)n^{-9/4} = o(1)$ to the drift $\mathrm{E}[Z_t - Z_{t+1} \mid X_t]$. Therefore, the assumption can reduce the drift by only $o(1)$.

Using Lemma 7, we define $h^*(z) := C \min\{\sqrt{\mu^2 - z}, \mu - \sqrt{\mu^2 - z}\}$ for $z < \mu$ such that $\mathrm{E}[Z_t - Z_{t+1} \mid X_t] \geq h^*(Z_t)$ for some sufficiently small constant $C > 0$ and all $0 < Z_t < \mu^2$. Moreover, we define $h^*(\mu^2) := h^*(\mu^2 - 1)$ (which sets the transition probability from state 0 of the X_t-process to that of state 1 as required). Note that Lemma 7 does not yield a bound on the drift if $X_t = \mu$; however, this is already our target state.

Lemma 9 shows the existence of a constant $c > 1$ such that $\frac{\mathrm{E}[(Z_t - Z_{t+1}) \cdot \mathbb{1}\{Z_{t+1} < Z_t\} | X_t]}{\mathrm{E}[(Z_{t+1} - Z_t) \cdot \mathbb{1}\{Z_{t+1} > Z_t\} | X_t]} \geq 2c^2$ for n large enough. With the $d(x)$ and $h^*(z)$ defined above, the fourth condition of Theorem 1 cannot be satisfied yet. Therefore, we replace $h^*(z)$ by an even smaller bound on the drift as follows. Assuming $c \leq 2$, we let

$$h(z) := \begin{cases} 0 & \text{if } z = 0, \\ C - \frac{C}{\ell} + \frac{C(c-1)\min\{\sqrt{\mu^2 - z}, \mu - \sqrt{\mu^2 - z}\}}{\ell} & \text{if } 0 < z < \mu^2, \\ h(\mu^2 - 1) & \text{if } z = \mu^2, \end{cases}$$

where C is the constant from the definition of h^*. It is easy to verify that $h^*(z) \geq h(z)$. Since $|X_{t+1} - X_t| \leq d(X_t)$, we get $h(Z_{t+1}) \leq h(Z_t)$ if $Z_{t+1} \leq Z_t \leq 3\mu^2/4$. Moreover, a simple case analysis depending on whether $Z_t \geq d^*/2$ or not and exploiting that $\mu \geq d^*$ proves that $h(Z_{t+1}) \leq ch(Z_t)$ if $Z_t \geq 3\mu^2/4$ and n large enough. Hence, we have satisfied the four conditions in Theorem 1. The minimum positive Z-value is $z_{\min} := \mu^2 - (\mu - 1)^2 = 2\mu - 1$ and we use the bound $h(z) \geq \frac{C(c-1)\min\{\sqrt{\mu^2 - z}, \mu - \sqrt{\mu^2 - z}\}}{\ell}$ for $z > 0$. If $Z_0 = 0$ or $Z_0 = \mu^2$, nothing is to show. Altogether, we get from the drift theorem that

$$\begin{aligned} \mathrm{E}[T \mid Z_0] &\leq 2c \left(\frac{z_{\min}}{h(z_{\min})} + \int_{z_{\min}}^{Z_0} \frac{1}{h(z)} \, \mathrm{d}z \right) \\ &\leq O(1) \cdot \left(\frac{2\mu}{C(c-1)/\ell} + \int_{2\mu-1}^{3\mu^2/4} \frac{\ell/(C(c-1))}{\mu - \sqrt{\mu^2 - z}} \, \mathrm{d}z + \int_{3\mu^2/4}^{\mu^2-1} \frac{\ell/(C(c-1))}{\sqrt{\mu^2 - z}} \, \mathrm{d}z \right) \\ &= O(\mu \log^2 n) + O(\log^2 n) \cdot \left(\left[\mu \ln(z) + 2\sqrt{\mu^2 - z} + \mu \ln \left(\frac{\mu - \sqrt{\mu^2 - z}}{\mu + \sqrt{\mu^2 - z}} \right) \right]_{2\mu-1}^{3\mu^2/4} \right. \\ &\qquad \left. + \left[-2\sqrt{\mu^2 - z} \right]_{3\mu^2/4}^{\mu^2-1} \right). \end{aligned}$$

The first integral evaluates to $O(\mu \log \mu)$ and the second one to $O(\mu)$. Since $\mu = o(n)$, the first hitting time for either 0 or μ is altogether $O(\mu \log^3 n)$. □

Note that unlike the analysis in Oliveto and Witt (2012), the previous lemma does not need a bound on the bandwidth of the population (i. e., the difference of best and worst individual). This constitutes a significant improvement in the sense that it might be used independently in future analyses of the **SGA**.

The results we have worked out so far will allow to prove that the **SGA** on OneMax is inefficient if $\mu \leq n^{1/4-\varepsilon}$. Lemma 10 gives a bound on the time for a bit to converge, i. e., for all individuals to either have a one or a zero at the bit position. We would like to know when all bits are converged for the first time. This is finally answered by the following lemma, which is very similar to Lemma 10 in Oliveto and Witt (2012).

Lemma 11. *Assume $\mu \leq n^{1/4-\varepsilon}$. Then with probability at least $1 - 2^{-\Omega(n^{\varepsilon/8}/\log^3 n)}$, after $\mu n^{\varepsilon/8}$ generations all bit positions have been converged at least once.*

Proof. We prove the statement by applying Markov's inequality iteratively to the probability that one bit position has not converged to a value after a certain time phase and then using a union bound to extend the calculation to all the bits in the bit string. In expected time $\mathrm{E}[T] \leq c\mu \log^3 n$ for some constant c according to Lemma 10, all the individuals have the same bit value at least in one position. By Markov's inequality with probability lower than 1/2 they don't all have the same bit after $2c\mu \log^3 n$ steps. Hence, after $n^{\varepsilon/8}/(2c \log^3 n)$ phases of length $2c\mu \log^3 n$ each, the probability that such bit has not been converged is at most $2^{-\Omega(n^{\varepsilon/8}/\log^3 n)}$.

Finally, by the union bound the probability that not all n bits have been converged at least once in $\mu n^{\varepsilon/8}$ generations is at most $n \cdot 2^{-\Omega(n^{\varepsilon/8}/\log^3 n)} = 2^{-\Omega(n^{\varepsilon/8}/\log^3 n)}$. □

We can now apply the machinery from the previous work (Oliveto and Witt, 2012).

3.2. Low Diversity

As in Oliveto and Witt (2012), we denote by s the number of bit positions that are not converged, which means that both bit values are taken by individuals in the population. The number of non-converged positions is a simple measure of diversity; for simplicity s is also called *the diversity* hereinafter. We study this measure since crossover does not have any effect on bit positions that are converged. Lemma 11 contains a statement for a single bit. We need to prove that almost all positions will be converged at any time w. o. p.; in other words, the diversity s is bounded. Throughout this section, we assume that $\mu \leq n^{1/4-\varepsilon}$ for some constant $\varepsilon > 0$.

To actually bound the diversity, we consider time phases of $T = \mu n^{\varepsilon/8}$ generations. In the first phase, diversity will collapse, and this will be maintained for the following phases.

Lemma 12. *Consider the* **SGA** *at some generation t, where $t \geq \mu n^{\varepsilon/8}$ and $t \leq 2^{n^{\varepsilon/10}}$. Then $s = O(\mu^2 n^{\varepsilon/8})$ at generation t with probability $1 - 2^{-\Omega(n^{\varepsilon/10})}$.*

Proof. By Lemma 11, in the first $\mu n^{\varepsilon/8}$ generations, all bits have converged at least once with probability $1 - 2^{-\Omega(n^{\varepsilon/9})}$. Now we consider an upper bound on the number of bits that have left the converged state by the end of the phase.

We define an indicator random variable $X_{i,j,k}$ for the event that the converged state of bit k is left when creating the j-th individual in the i-th generation of the phase, where $1 \leq i \leq \mu n^{\varepsilon/8}$, $1 \leq j \leq \mu$ and $1 \leq k \leq n$.

To leave the converged state, the bit position must be flipped at least once. Since each bit flips with probability $1/n$, we get $\text{Prob}[X_{i,j,k} = 1] = 1/n$, and the expected value of the sum S of the $X_{i,j,k}$ is

$$\text{E}[S] = \sum_i \sum_j \sum_k \text{Prob}[X_{i,j,k}] = \frac{\mu \cdot T \cdot n}{n} = \mu \cdot T = \mu^2 n^{\varepsilon/8}.$$

Obviously, S is an upper bound on the number of positions that leave the converged state. By Chernoff bounds $\text{E}[S] \leq 2\mu^2 n^{\varepsilon/8}$ with probability $1 - 2^{-\Omega(n^{\varepsilon/8})}$, which, together with the fact that all bits converge at least once in the phase proves the statement for generation $\mu n^{\varepsilon/8}$. For later generations, the statement follows by considering additional phases of length $\mu n^{\varepsilon/8}$. The total failure probability in at most $2^{n^{\varepsilon/10}}$ generations is still $2^{-\Omega(n^{\varepsilon/10})}$. □

Roughly speaking, for $\mu \leq n^{1/4-\varepsilon}$, this means that the number of non-converged positions is $O(n^{1/2-15\varepsilon/8})$ for an exponential number of generations. In more simple terms, we have $s = O(n^{1/2-\varepsilon})$, which will be essential in the following.

Finally, in the very first generations before bits have converged, we use the following rough estimate of the progress. The lemma is the same as Lemma 5 in Oliveto and Witt (2012).

Lemma 13. *With probability at least $1 - 2^{-\Omega(n^{2\varepsilon})}$ the maximum progress achieved by the crossover operator in one step is $n^{1/2+\varepsilon}$. With probability at least $1 - 2^{-\Omega(n^{2\varepsilon})}$ the maximum progress achieved by a crossover and mutation step is $2n^{1/2+\varepsilon}$. The maximum progress per generation is bounded in the same way if $\mu = poly(n)$.*

Proof. We consider the progress achieved in one crossover step (i. e., the number of one-bits gained compared to the current best individual). Let the best individual in the population have i one-bits and $n - i$ zero-bits. We pessimistically add one-bits so they both have i to the two individuals selected for crossover in a way that minimizes the overlapping one-bits (and consider $i \geq n/2$: the opposite case is symmetrical by considering zeroes instead of ones). Then we will have $2(n - i)$ positions with a one-bit and a zero-bit and $n - 2(n - i)$ positions with overlapping one-bits. Hence the expected number of one-bits in the offspring is

$$\frac{1}{2}2(n - i) + n - 2(n - i) = n - (n - i) = i$$

which implies no progress in expectation (i. e., $i - i = 0$). Hence, the progress depends on the variance which is maximized for $i = n/2$, where we get an expected number of $n/2$ one-bits in the offspring. By Chernoff bounds, in this case, the offspring will not have more than

$n/2 + n^{1/2+\varepsilon}$ one-bits with probability at least $1 - 2^{-\Omega(n^{2\varepsilon})}$. Hence, w. o. p. the maximum progress in each step is at most $n^{1/2+\varepsilon}$.

We now take mutation into account. The probability that $n^{1/2}$ bits are flipped is upper bounded by

$$\binom{n}{\sqrt{n}}\left(\frac{1}{n}\right)^{\sqrt{n}} \leq \frac{1}{\sqrt{n}!} = 2^{-\Omega(n^{1/2}\ln n)}.$$

By adding the failure probabilities and pessimistically assuming all the flipped bits create one-bits, we get a maximum progress after crossover *and* mutation that is bounded by $n^{1/2+\varepsilon} + n^{1/2} \leq 2n^{1/2+\varepsilon}$ with probability $1 - 2^{-\Omega(n^{\varepsilon})} - 2^{-\Omega(n^{1/2}\ln n)} = 1 - 2^{-\Omega(n^{\varepsilon})}$ (assuming ε small enough).

Taking a union bound over $\mu = \mathit{poly}(n)$ steps per generation, the maximum progress per generation is greater than $2n^{1/2+\varepsilon}$ with probability $\mu 2^{-\Omega(n^{\varepsilon})} = 2^{-\Omega(n^{\varepsilon})}$. □

3.3. Drift of Best and Worst Fitness Values

In this section, the whole population is proved to drift towards the center of the boolean hypercube, which results in exponential optimization time. In the analysis, we consider the quantities h (the best OneMax-value of a population), and ℓ (the worst OneMax-value). Obviously, if $h < n$ then the optimum has not been found. The rest of this section is, up to minor adjustements reflecting the new bound on μ, basically a copy of Section 3.2 in the journal version of Oliveto and Witt (2012).

The aim is to bound h and ℓ in a drift analysis using a so-called potential function. Similarly as in Neumann et al. (2009), the potential of an individual x is defined by $g(x) := e^{\kappa \textsc{OneMax}(x)}$ for some $\kappa := \kappa(n)$ to be chosen later, and $g(X) := \sum_{i=1}^{\mu} g(x_i)$ for every population $X := \{x_1, \ldots, x_\mu\}$ (note that populations are multisets). Let us consider a current population X_t (note that X_t has been redefined) at generation t and the process of creating the next population X_{t+1} at generation $t+1$ (dropping the time indices unless there is risk of confusion). This process consists of μ consecutive operations choosing two parent individuals, crossing them over and mutating the result. Let P_i and Q_i be the two random parent individuals in the i-th operation (at generation t), $1 \leq i \leq \mu$, and let K_i be the random offspring. The next lemma notes an important observation on the OneMax-value of the offspring.

Hereinafter, $\Delta^{(\mathrm{m})}(j)$ denotes the random change in OneMax-value when applying standard bit mutation to an individual with j one-bits, $\mathrm{Bin}(a, b)$ still denotes a random variable following the binomial distribution with parameters a and b, and $H(\cdot, \cdot)$ denotes the Hamming distance.

Lemma 14. *It holds that*

$$|K_i| = \frac{|P_i| + |Q_i| + 2C(P_i, Q_i)}{2} + \Delta^{(\mathrm{m})}(|P_i|/2 + |Q_i|/2 + C(P_i, Q_i)),$$

where $C(P_i, Q_i) = \mathrm{Bin}(H(P_i, Q_i), 1/2) - H(P_i, Q_i)/2$. *Moreover,*

$$|K_i| = \frac{|P_i| + C(P_i, Q_i) + 2\Delta^*(|P_i| + C(P_i, Q_i))}{2}$$

$$+\frac{|Q_i|+C(P_i,Q_i)+2\Delta^*(|Q_i|+C(P_i,Q_i))}{2},$$

where $\Delta^*(j) := \mathrm{Bin}(n/2-j/2,1/n)-\mathrm{Bin}(j/2,1/n)$ *is the random increase in one-bits given that each bit in a string of length* $n/2$ *with* $j/2$ *one-bits is flipped with probability* $1/n$*, i. e., half the standard mutation probability.*

Proof. By definition, the crossover part of the i-th operation leads to $|P_i\cap Q_i|+\mathrm{Bin}(H(P_i,Q_i),1/2)$ one-bits before mutation. Moreover $|P_i\cup Q_i| = |P_i\cap Q_i| + H(P_i,Q_i)$, which means that $(1/2)(|P_i|+|Q_i|) = |P_i\cap Q_i|+H(P_i,Q_i)/2$. Therefore, an individual with $(1/2)(|P_i|+|Q_i|)+C(P_i,Q_i)$ one-bits is subjected to mutation, which is the first statement of the lemma. The increase in one-bits due to mutation is a random variable with distribution

$$\begin{aligned}&\mathrm{Bin}(n-(|P_i|/2+|Q_i|/2+C(P_i,Q_i)),1/n)-\mathrm{Bin}(|P_i|/2+|Q_i|/2+C(P_i,Q_i),1/n)\\&=\mathrm{Bin}(n/2-|P_i|/2-C(P_i,Q_i)/2,1/n)+\mathrm{Bin}(n/2-|Q_i|/2-C(P_i,Q_i)/2,1/n)\\&\quad-\Big(\mathrm{Bin}(|P_i|/2+C(P_i,Q_i)/2,1/n)+\mathrm{Bin}(|Q_i|/2+C(P_i,Q_i)/2,1/n)\Big),\end{aligned}$$

where the equality follows from the fact that if $X_1=\mathrm{Bin}(n_1,p)$ and $X_2=\mathrm{Bin}(n_2,p)$ then $X_1+X_2=\mathrm{Bin}(n_1+n_2,p)$. The second statement follows now by regrouping terms. □

Due to linearity of expectation and $\mathrm{E}[C(P_i,Q_i)]=0$, we have $\mathrm{E}[\Delta^*(|P_i|+C(P_i,Q_i))] = 1/2-|P_i|/n$, and analogously for Q_i. This results in $\mathrm{E}[|K_i|] = (|P_i|/2+(1/2-|P_i|/n))+(|Q_i|/2+(1/2-|Q_i|/n))$. In other words, the random K_i depends on the random P_i and Q_i, whereas $\mathrm{E}[|K_i|]$ only depends on $|P_i|$ and $|Q_i|$, each of which has weight $1/2$. Considering $\mathrm{E}[|K_i|]$, we see that one operation is "split" into two analogous terms, whose values are determined by $|P_i|$ and $|Q_i|$, respectively.

However, the random potential of the offspring is given by $e^{\kappa|K_i|}$, and we have to bound $\mathrm{E}[e^{\kappa|K_i|}]$. We will see below that $\mathrm{E}[e^{\kappa|K_i|}]$ is not too different from $e^{\mathrm{E}[|K_i|]}$ for small κ. Actually, we will also "split" an operation into two terms that "basically" only depend on P_i and Q_i, respectively.

Lemma 15.

$$e^{\kappa|K_i|}\le\frac{e^{\kappa(|P_i|+C(P_i,Q_i)+2\Delta^*(|P_i|+C(P_i,Q_i)))}}{2}+\frac{e^{\kappa(|Q_i|+C(P_i,Q_i)+2\Delta^*(|Q_i|+C(P_i,Q_i)))}}{2}.$$

Proof. The statement follows from Lemma 14 since the geometric mean is at most the arithmetic mean, i. e., $e^{a/2}\cdot e^{b/2}\le e^a/2+e^b/2$ for arbitrary a and b. □

Both terms on the right-hand side have in common that they depend on the random $C(P_i,Q_i)$. The influence of $C(P_i,Q_i)$ will be neglegible for sufficiently small κ, as the following lemma shows.

Lemma 16. *Let* $s=H(P_i,Q_i)\ge 1$*. If* $|P_i|\ge(1+c)(n/2)$ *for some arbitrarily small constant* $c>0$ *and* $s\le(c/4)n$ *then choosing* $\kappa\le\frac{c}{2500s}$ *yields*

$$\mathrm{E}[e^{\kappa(C(P_i,Q_i)+2\Delta^*(|P_i|+C(P_i,Q_i)))}]\le 1-c_1\kappa$$

for some constant $c_1 > 0$. If the assumption on $|P_i|$ is dropped and c is small enough then

$$\mathrm{E}[e^{\kappa(C(P_i,Q_i)+2\Delta^*(|P_i|+C(P_i,Q_i)))}] \leq 1 + c_2\kappa$$

for some constant $c_2 > 0$.

Proof. We abbreviate $\Psi(P_i, Q_i) := e^{\kappa(C(P_i,Q_i)+2\Delta^*(|P_i|+C(P_i,Q_i)))}$. This random variable, whose expectation has to be bounded, is dependent on the combined effect of crossover and mutation. Note that Δ^* is decreasing in its argument and that we have $C(P_i, Q_i) \geq -s$. In the following, we work with the upper bound

$$\Psi(P_i, Q_i) \leq e^{\kappa C(P_i,Q_i)} \cdot e^{2\kappa\Delta^*(|P_i|-s)}$$

and assume that $|P_i| \geq (1 + c)(n/2)$. Since crossover and mutation work independently of each other, we get

$$\mathrm{E}[\Psi(P_i, Q_i)] \leq \mathrm{E}[e^{\kappa C(P_i,Q_i)}] \cdot \mathrm{E}[e^{2\kappa\Delta^*(|P_i|-s)}].$$

We concentrate first on the first expectation. By definition, $C(P_i, Q_i) = \mathrm{Bin}(s, 1/2) - s/2$. Using the moment-generating function of the binomial distribution, we obtain

$$\begin{aligned}\mathrm{E}[e^{\kappa C(P_i,Q_i)}] &= \mathrm{E}[e^{\kappa(-s/2)} \cdot e^{\kappa\mathrm{Bin}(s,1/2)}] \\ &= e^{-\kappa s/2} \cdot \left(\frac{1}{2} + \frac{1}{2}e^{\kappa}\right)^s.\end{aligned}$$

Assuming that $\kappa \leq 1$, we use the inequality $e^x \leq 1 + x + x^2$ for $x \leq 1$ and obtain

$$\begin{aligned}\mathrm{E}[e^{\kappa C(P_i,Q_i)}] &\leq e^{-\kappa s/2} \cdot \left(1 + \frac{\kappa}{2} + \kappa^2\right)^s \leq e^{-\kappa s/2} e^{(\kappa/2+\kappa^2)s} \\ &= e^{\kappa^2 s},\end{aligned}$$

which for $\kappa = c/(2500s)$ (a choice that will turn out useful later) gives us the upper bound

$$\mathrm{E}[e^{\kappa C(P_i,Q_i)}] \leq e^{c^2/(2500^2 s)} \leq 1 + \frac{c^2}{3125000s}$$

using $e^x \leq 1 + 2x$ for $x \leq 1$ and $c^2/(2500^2 s) \leq c \leq 1$.

Next we deal with the effect of mutation, more precisely we bound the expected value of $e^{2\kappa\Delta^*(|P_i|-s)}$. Following the proof of the simplified drift theorem Oliveto and Witt (2011), we first bound the plain drift $\mathrm{E}[\Delta^*(|P_i| - s)]$ and then its moment-generating function. Recall that $\Delta^*(j) = \mathrm{Bin}(n/2 - j/2, 1/n) - \mathrm{Bin}(j/2, 1/n)$ is the random increase in OneMax-value when a bit string of length $n/2$, containing $j/2$ ones, is subject to standard bit mutation with probability $1/n$. Hence, we have

$$\mathrm{E}[\Delta^*(|P_i| - s)] = \frac{1}{2} - \frac{|P_i| - s}{n} \leq \frac{-c}{4},$$

where the last inequality follows by the assumptions made in the lemma.

Moreover, we know that the number of flipping bits follows an exponential decay, more precisely

$$\text{Prob}[\Delta^*(i) = z] \le \binom{n/2}{|z|} \left(\frac{1}{n}\right)^{|z|} \le \frac{1}{|z|!} \le e^{-|z|+2}$$

for any i and any $z \in \mathbb{Z}$. We get

$$\begin{aligned}\mathrm{E}[e^{|\Delta^*(i)|/2}] &= \sum_{z\ge 0} e^{z/2} \,\text{Prob}[|\Delta^*(i)| = z] = \sum_{z\in\mathbb{Z}} e^{|z|/2} \,\text{Prob}[\Delta^*(i) = z] \\ &\le 2\sum_{z\ge 0} e^{z/2} e^{-|z|+2} \le 2\sum_{z\ge 0} e^{2-z/2} = \frac{2e^2}{1-e^{-1/2}} < 38.\end{aligned}$$

Expanding the moment-generating function $\mathrm{E}[e^{\lambda\Delta^*(i)}]$, this implies for any $\lambda \le 1/2$ that

$$\begin{aligned}\mathrm{E}[e^{\lambda\Delta^*(i)}] &= 1 + \lambda\,\mathrm{E}[\Delta^*(i)] + \sum_{z=2}^{\infty} \frac{\mathrm{E}[(\lambda\Delta^*(i))^z]}{z!} \le 1 + \lambda\,\mathrm{E}[\Delta^*(i)] + \sum_{z=2}^{\infty} \frac{\mathrm{E}[(\lambda|\Delta^*(i)|)^z]}{z!} \\ &\le 1 + \lambda\,\mathrm{E}[\Delta^*(i)] + \lambda^2 \sum_{z=0}^{\infty} \frac{\frac{\mathrm{E}[(|\Delta^*(i)|/2)^z]}{z!}}{(1/2)^2} \\ &\le 1 + \lambda\,\mathrm{E}[\Delta^*(i)] + \frac{\lambda^2}{1/4}\,\mathrm{E}[e^{|\Delta^*(i)|/2}] \\ &\le 1 + \lambda\,\mathrm{E}[\Delta^*(i)] + 152\lambda^2.\end{aligned}$$

Identifying $\lambda = 2\kappa$ and assuming $2\kappa \le -\mathrm{E}[\Delta^*(i)]/304$, a simple upper bound is obtained from this as follows:

$$\mathrm{E}[e^{2\kappa\Delta^*(i)}] \le 1 + 2\kappa\,\mathrm{E}[\Delta^*(i)] - \frac{1}{2}(2\kappa)\,\mathrm{E}[\Delta^*(i)] \le 1 + \kappa\,\mathrm{E}[\Delta^*(i)].$$

Since $\mathrm{E}[\Delta^*(i)] \le -c/4$ and thus $-\mathrm{E}[\Delta^*(i)]/304 \ge c/1216$, the choice $\kappa := c/(2500s)$ from above satisfies the condition $2\kappa \le -\mathrm{E}[\Delta^*(i)]/304$ already for $s \ge 1$, and we get

$$\mathrm{E}[e^{2\kappa\Delta^*(i)}] \le 1 - \frac{c^2}{10000s}.$$

Altogether, the random variable under consideration has been bounded according to

$$\Psi(P_i, Q_i) \le \left(1 + \frac{c^2}{3125000s}\right)\left(1 - \frac{c^2}{10000s}\right) \le 1 - \frac{c^2}{20000s},$$

which is $1 - c_1\kappa$ for some constant $c_1 > 0$.

If the assumption on $|P_i|$ is dropped, then we work with the trivial bound $\mathrm{E}[\Delta^*(i)] \le 1$. Recalling that $\mathrm{E}[e^{\lambda\Delta^*(i)}] \le 1 + \lambda\,\mathrm{E}[\Delta^*(i)] + 152\lambda^2$, we get

$$\mathrm{E}[e^{2\kappa\Delta^*(i)}] \le 1 + \frac{2c}{2500s} + 152 \cdot \frac{4c^2}{(2500s)^2} \le 1 + \frac{610c}{2500s}$$

as $c \leq 1$ and $s \geq 1$. Then (for small enough c)

$$\mathrm{E}[\Psi(P_i, Q_i)] \leq \left(1 + \frac{c^2}{3125000s}\right)\left(1 + \frac{610c}{2500s}\right) \leq 1 + \frac{611c}{2500s},$$

which is $1 + c_2\kappa$ for another constant $c_2 > 0$. □

Our aim is to bound $\mathrm{E}[g(X_{t+1}) - g(X_t) \mid X_t]$. Let S_i denote the number of times that individual $x_i := x_i^{(t)}$ is chosen as first or second parent in a crossover operation during the μ operations. We consider the other random parent $y_i(j)$ in the j-th operation choosing x_i and study the potential of their random offspring $K(x_i, y_i(j))$ (created by a crossover-mutation sequence). We get

$$g(X_{t+1}) = \frac{1}{2}\sum_{i=1}^{\mu}\sum_{j=1}^{S_i} e^{\kappa|K(x_i,y_i(j))|},$$

where the factor $1/2$ accounts for the fact that the double sum counts each offspring twice, namely once for each parent. Using Lemma 15, we get

$$\begin{aligned}
&g(X_{t+1}) \\
&\leq \frac{1}{2}\sum_{i=1}^{\mu}\sum_{j=1}^{S_i}\left(\frac{e^{\kappa(|x_i|+C(x_i,y_i(j))+2\Delta^*(|x_i|+C(x_i,y_i(j))))}}{2} + \frac{e^{\kappa(|y_i(j)|+C(x_i,y_i(j))+2\Delta^*(|y_i(j)|+C(x_i,y_i(j))))}}{2}\right) \\
&= \sum_{i=1}^{\mu}\sum_{j=1}^{S_i}\frac{e^{\kappa(|x_i|+C(x_i,y_i(j))+2\Delta^*(|x_i|+C(x_i,y_i(j))))}}{2},
\end{aligned}$$

where the equality holds since each second term in the big parentheses also appears as first term for another index i, more precisely when i indexes the individual called "$y_i(j)$".

Now let x_i' denote the worst case from the random $y_i(j)$, more precisely the one that makes the offspring potential stochastically largest. Since crossover and mutation work independently of selection, we get the following bound on the potential of the offspring population:

$$\mathrm{E}[g(X_{t+1}) \mid X_t] \leq \sum_{i=1}^{\mu}\mathrm{E}[S_i] \cdot \frac{\mathrm{E}[e^{\kappa(|x_i|+C(x_i,x_i')+2\Delta^*(|x_i|+C(x_i,x_i')))}]}{2}. \tag{2}$$

The following simple lemma bounds $\mathrm{E}[S_i]$.

Lemma 17. $\mathrm{E}[S_i] \leq 2h/\ell$.

Proof. The probability of choosing a given individual is maximized if its value is h and the population consists of $\mu - 1$ individuals with value ℓ and one of value h. Hence, the probability that the individual is chosen as parent is at most $h/((\mu-1)\ell + h) \leq h/(\mu\ell)$. Since 2μ parents are chosen, the lemma follows from the linearity of expectation. □

More effort is needed to bound the second expectation in the right-hand side of (2). If $g(X_t)$ is large, then the following lemma will help us to obtain a negative drift. It assumes $s \le n^{1/2-\varepsilon}$ (which follows from Lemma 12 for $\mu \le n^{1/4-\varepsilon}$ and sufficiently large n).

Lemma 18. *Suppose that $s \le n^{1/2-\varepsilon}$, $\kappa := c/(2500s)$ for some constant $c > 0$ and $\mu = poly(n)$. If $g(X) \ge e^{\kappa(1+2c)n/2}$, then there is a non-empty set $X^* \subset X$ of individuals $x \in X$ satisfying $|x| \ge (1+c)n/2$. Moreover, $g(X) = (1 + 2^{-\Omega(n^{1/2+\varepsilon})}) \sum_{x \in X^*} g(x)$.*

Proof. Asssume $X^* = \emptyset$. Then $g(X) \le \mu e^{\kappa(1+c)n/2} = e^{\kappa(1+c)n/2 + \ln \mu}$. Since $\kappa = c/(2500s) = \Omega(n^{-1/2+\varepsilon})$ and $\ln \mu = O(\log n)$ by our assumption, we arrive at the contradiction $\kappa(1+c)n/2 + \ln \mu \le \kappa(1+1.5c)n/2$ if n is not too small. The second claim follows since $\kappa(1+1.5c)n/2 = \kappa(1+2c)n/2 - \Omega(n^{1/2+\varepsilon})$. □

The next lemma states a multiplicative drift of the potential away from large values, assuming some minimum value of the worst individual.

Lemma 19. *If $s \le n^{1/2-\varepsilon}$, $\kappa := c/(2500s)$ for some constant $c > 0$, $\mu = poly(n)$ and $g(X_t) \ge e^{\kappa(1+2c)n/2}$, then*

$$\mathrm{E}[g(X_{t+1}) \mid X_t] \le (1 - c_3\kappa) \cdot g(X_t).$$

for some constant $c_3 > 0$.

Proof. According to Lemma 18, there is a subset $X^* \subset X_t$ such that

$$g(X_t) = \sum_{x \in X^*} g(x) + 2^{-\Omega(n^{1/2+\varepsilon})} \sum_{x \in X^*} g(x)$$

We have already argued that

$$\mathrm{E}[g(X_{t+1}) \mid X_t)] \le \sum_{i=1}^{\mu} \mathrm{E}[S_i] \cdot \frac{\mathrm{E}[e^{\kappa(|x_i| + C(x_i,x_i') + 2\Delta^*(|x_i| + C(x_i,x_i')))}]}{2}.$$

By Lemma 17, $\mathrm{E}[S_i] \le 2h/\ell \le 2(\ell+s)/\ell$. Since $\ell \ge n/10$, we get $\mathrm{E}[S_i] = 2 + O(s/n) = 2 + O(n^{-1/2-\varepsilon})$. Hence, for the i such that $x_i \in X^*$ we get from Lemma 16 that

$$\mathrm{E}[S_i] \cdot \frac{\mathrm{E}[e^{\kappa(|x_i| + C(x_i,x_i') + 2\Delta^*(|x_i| + C(x_i,x_i')))}]}{2} \le (1 + O(n^{-1/2-\varepsilon}))(1 - c_1\kappa)e^{\kappa|x_i|} = (1 - c_4\kappa)e^{\kappa|x_i|}$$

for some constant $c_4 > 0$, using $\kappa = c/(2500s) = \Omega(n^{-1/2+\varepsilon})$. (At this place, our bound on s is crucial.) For the $x_i \notin X^*$ we know by Lemma 16 that

$$\mathrm{E}[S_i] \cdot \frac{\mathrm{E}[e^{\kappa(|x_i| + C(x_i,x_i') + 2\Delta^*(|x_i| + C(x_i,x_i')))}]}{2} \le (1 + O(n^{-1/2-\varepsilon}))(1 + c_2\kappa)e^{\kappa|x_i|} = (1 + c_5\kappa)e^{\kappa|x_i|}$$

for some constant $c_5 > 0$. Altogether,

$$\begin{aligned}
\mathrm{E}[g(X_{t+1}) \mid X_t] &\leq (1 - c_4\kappa) \left(\sum_{x \in X^*} g(x) \right) + 2^{-\Omega(n^{1/2+\varepsilon})}(1 + c_5\kappa) \sum_{x \notin X^*} g(x) \\
&\leq (1 - c_4\kappa) \left(\sum_{x \in X_t} g(x) \right) + 2^{-\Omega(n^{1/2+\varepsilon})}(1 + c_5\kappa) \sum_{x \in X_t} g(x) \\
&\leq \left(1 - c_4\kappa + 2^{-\Omega(n^{1/2+\varepsilon})}\right) g(X_t) = (1 - c_3\kappa) g(X_t)
\end{aligned}$$

for some constant $c_3 > 0$. □

We will apply the following simplified negative-drift theorem (proved in Appendix A) with slightly weaker third condition than in Oliveto and Witt (2012).

Theorem 2 (Simplified Drift with Scaling 2013)**.** *Let X_t, $t \geq 0$, be real-valued random variables describing a stochastic process over some state space. Suppose there exist an interval $[a, b] \subseteq \mathbb{R}$ and, possibly depending on $\ell := b - a$, a drift bound $\varepsilon := \varepsilon(\ell) > 0$ as well as a scaling factor $r := r(\ell)$ such that for all $t \geq 0$ the following three conditions hold:*

1. $\mathrm{E}[X_{t+1} - X_t \mid X_0, \ldots, X_t \,;\, a < X_t < b] \geq \varepsilon$,
2. $\mathrm{Prob}[|X_{t+1} - X_t| \geq jr \mid X_0, \ldots, X_t \,;\, a < X_t] \leq e^{-j}$ *for* $j \in \mathbb{N}_0$,
3. $1 \leq r^2 \leq \varepsilon\ell/(132 \log(r/\varepsilon))$.

Then for the first hitting time $T^* := \min\{t \geq 0 \colon X_t \leq a \mid X_0, \ldots, X_t \,;\, X_0 \geq b\}$ it holds that $\mathrm{Prob}[T^* \leq e^{\varepsilon\ell/(132r^2)}] = O(e^{-\varepsilon\ell/(132r^2)})$.

The bound from Lemma 19 results in the following lemma, which says that the population is "centered" in the middle of the hypercube for an exponential number of generations, assuming that the diversity is bounded.

Lemma 20. *Assuming $s \leq n^{1/2-\varepsilon}$ the whole time, all populations up to generation $2^{c'n^\varepsilon}$, for some constant $c' > 0$, satisfy $\ell \geq (1-c)n/2$ and $h \leq (1+c)n/2$ with probability $1 - 2^{-\Omega(n^\varepsilon)}$, where $c > 0$ is an arbitrarily small constant.*

Proof. Lemma 19 states a multiplicative drift but Theorem 2 is for an additive setting. Hence, as in Neumann et al. (2009), we switch over to the potential function $g'(X) := \ln g(X)$. Since ln is concave, Jensen's inequality yields

$$\mathrm{E}[g'(X_{t+1}) \mid X_t] = \mathrm{E}[\ln(g(X_{t+1})) \mid X_t] \leq \ln(\mathrm{E}[g(X_{t+1}) \mid X_t]).$$

Hence, if $g'(X_t) \geq \kappa(1 + 2c)n/2$ then by Lemma 19

$$\begin{aligned}
&\mathrm{E}[g'(X_{t+1}) \mid X_t] \leq \ln((1 - \Omega(\kappa))g(X_t)) \\
&= \ln(1 - \Omega(\kappa)) + \ln(g(X_t)) = -\Omega(\kappa) + g'(X_t),
\end{aligned}$$

which establishes the additive drift $\mathrm{E}[g'(X_{t+1}) - g'(X_t) \mid X_t] = -\Omega(\kappa)$. We now switch to the potential function $g''(X) := \kappa(1+4c)n/2 - g'(X)$, where the negation is necessary to fit the perspective of Theorem 2, and the drift interval $a := 0, b := \kappa(1+4c)n/2 - \kappa(1+2c)n/2 = \kappa cn$. Formally, the random variables called X_t in the drift theorem are identified with the stochastic process $g''(X_t)$ considered here.

We obtain a drift of $\Omega(\kappa)$ for all $g''(X)$ such that $a \leq g''(X) \leq b$. The first condition of Theorem 2 has been established with $\varepsilon_{\text{drift}} = \Omega(\kappa)$ (where $\varepsilon_{\text{drift}}$ denotes the parameter called ε in the drift theorem). Moreover, by Chernoff bounds $g''(X_0) \geq b$ at initialization of the **SGA** with probability $1 - 2^{-\Omega(n^{1/2})}$. We start the drift analysis at the first point of time T^* where diversity has collapsed (formally, time 0 in the drift theorem corresponds to time T^*). By our assumptions, $T^* \leq \mu n^{\varepsilon/8}$. Using Lemma 13, we still have $g''(X_{\mu n^{\varepsilon/8}}) \geq b$, i. e., the stochastic process considered in the drift theorem starts above b as required.

To prove the second condition of the drift theorem, we note that

$$g'_{\min} := \ln \mu + \kappa\ell \;\leq\; g'(X_t) \;\leq\; \ln \mu + \kappa h =: g'_{\max}$$

and

$$g'_{\max} - g'_{\min} \leq \frac{c}{2500s}(h - \ell) \leq \frac{c}{2500}$$

as $h - \ell \leq s$. Hence, also the potential $g''(X)$ will not change by more than $c/2500$ if all individuals are replaced by the best or worst individual in order to maximize the probability of jumping towards or away from the optimum. We set $r := c/2500 + 2\kappa \max\{s, n^{\varepsilon/2}\} = O(n^{\varepsilon/2})$. Crossover can change the potential with respect to either parent by at most $s \leq r/2$. To change the potential by jr, where $j \geq 1$, at least $(jr/\kappa - r/2) \geq (j - 1/2)r/\kappa \geq (j - 1/2)n^{\varepsilon/2}$ bits have to flip in at least one of the $\mu = \mathit{poly}(n)$ mutations that happen in a generation. This probability is easily bounded by e^{-j} if n is large enough. This verifies the second condition. Altogether, the parameters of the drift theorem satisfy $\ell = b - a = \Omega(n^{1/2+\varepsilon})$, $\varepsilon_{\text{drift}} = \Omega(n^{-1/2+\varepsilon})$ and $r = O(n^{\varepsilon/2})$. The third condition of the drift theorem is now easily verified since $\varepsilon_{\text{drift}}\ell/\log(r/\varepsilon_{\text{drift}}) = \Omega(n^{2\varepsilon}/\log n)$. Since $\varepsilon_{\text{drift}}\ell/r^2 = \Omega(n^{\varepsilon})$, the time to pass the drift interval is $2^{\Omega(n^{\varepsilon})}$ w. o. p. If $g(X) \leq e^{\kappa(1+4c)n/2}$ by definition no individual from X can have more than $(1+4c)n/2$ one-bits, in particular the optimum is not reached.

A symmetrical argument can be applied to the minimum OneMax-value of the individuals in the population. □

As proved in the previous subsection, we indeed have $s \leq n^{1/2-\varepsilon}$ for an exponential number of generations w. o. p. In each generation there is a probability of only $2^{-\Omega(n^{\varepsilon})} + 2^{-\Omega(n^{\varepsilon/10})}$ (the latter error term stems from Lemma 12) that one of our assumptions (low diversity and $\ell \geq n/10$) is not satisfied. Since the sum of the failure probabilities within $2^{n^{\varepsilon/11}}$ generations is still $2^{-\Omega(n^{\varepsilon/10})}$, we have proved the following main result.

Theorem 3. *Let $\mu \leq n^{1/4-\varepsilon}$ for an arbitrarily small constant $\varepsilon > 0$. Then with probability $1 - 2^{-\Omega(n^{\varepsilon/10})}$, the* **SGA** *on* OneMax *does not create individuals with more than $(1+c)n/2$ or less than $(1-c)n/2$ one-bits, where $c > 0$ is an arbitrarily small constant, within the first $2^{n^{\varepsilon/11}}$ generations. In particular it does not reach the optimum then.*

Figure 2: Converged bits vs number of generations and $\mu = (1/6)\sqrt{n}, \sqrt{n}, 3\sqrt{n}, 6\sqrt{n}$ for $n = 2^{12}$.

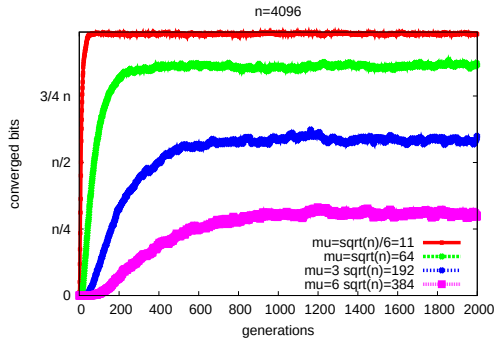
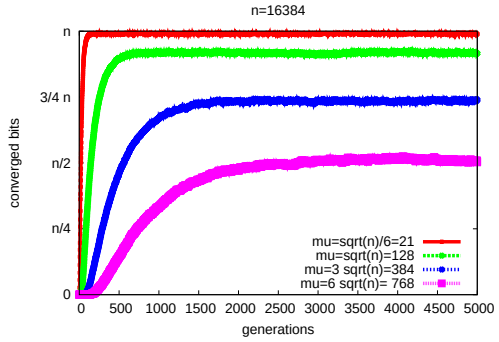


Figure 3: Converged bits vs number of generations and $\mu = (1/6)\sqrt{n}, \sqrt{n}, 3\sqrt{n}, 6\sqrt{n}$ for $n = 2^{14}$.



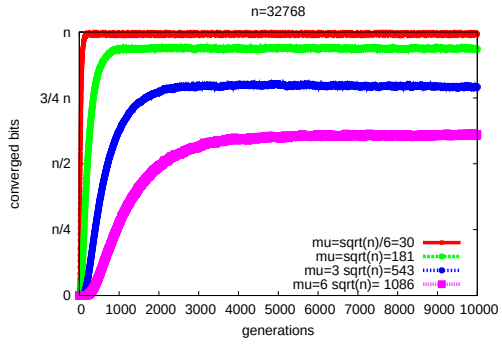
4. Discussion, Experiments and Conclusions

An improved analysis of the **SGA** for OneMax has been presented. Through new combinations and extensions of state-of-the art techniques for the analysis we have bounded the diversity of the population up to $\mu \leq n^{1/4-\varepsilon}$ without requiring a bound on its bandwidth. This generality will very likely allow to re-use the new techniques in future analyses of the **SGA**.

Our analysis requires that, at each generation, a large number of the bits in the population are converged. In particular, Lemma 12 bounds the fraction of non-converged bits by $o(1)$ only if $\mu = o(n^{1/2})$. We only managed to show a result for $\mu = O(n^{1/4-\varepsilon})$ since the proof of Lemma 19 requires $\kappa = \omega(s/n)$, while $\kappa = O(1/s)$. This means that $s = o(n^{1/2})$ is required, and by Lemma 12 we get $\mu = O(\sqrt{s}) = o(n^{1/4})$. It is an open problem to remove the requirement $\kappa = \omega(s/n)$.

We perform some preliminary experiments to further look into the number of converged bits during runs of the algorithm. We consider exponentially growing problem sizes n, concentrate the population size around $\mu = c \cdot \sqrt{n}$ with $c = 1/6, 1, 3, 6$ and plot the number of converged bits at each generation. The figures show how, for the considered problem sizes, the number of converged bits indeed drops around $\mu = O(n^{1/2})$. In particular, for

Figure 4: Converged bits vs number of generations and $\mu = (1/6)\sqrt{n}, \sqrt{n}, 3\sqrt{n}, 6\sqrt{n}$ for $n = 2^{15}$.



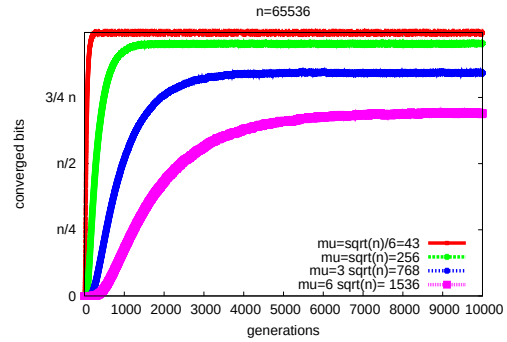


Figure 5: Converged bits vs number of generations and $\mu = (1/6)\sqrt{n}, \sqrt{n}, 3\sqrt{n}, 6\sqrt{n}$ for $n = 2^{16}$.

$\mu = \sqrt{n}/6$ the number of converged bits would be sufficient for Lemma 12 to hold, while for $c \geq 1$ the number of converged bits would not be sufficient. Nevertheless as n grows we can also see that, if the time for bits to converge after initialization increases quickly, also the number of converged bits increases for a fixed population size μ. Hence, we cannot rule out the possibility that for sufficiently large n a sufficient number of bits converge unless we perform more extensive experiments with more computation power which we leave for future work. In any case, if $\mu = \omega(n^{1/2})$, completely different techniques might be needed.

References

Auger, A., Doerr, B. (Eds.), 2011. Theory of Randomized Seach Heuristics–Foundations and Recent Developments. World Scientific.

Doerr, B., Doerr, C., Ebel, F., 2013. Lessons from the black-box: Fast crossover-based genetic algorithms, in: Proc. of GECCO '13, ACM Press. pp. 781–788.

Doerr, B., Johannsen, D., Kötzing, T., Neumann, F., Theile, M., 2010. More effective crossover operators for the all-pairs shortest path problem, in: Proc. of PPSN '10, Springer. pp. 184–193.

Feldmann, M., Kötzing, T., 2013. Optimizing expected path lengths with ant colony optimization using fitness proportional update, in: Proc. of FOGA '13, ACM Press. pp. 65–74.

Feller, W., 1971. An Introduction to Probability Theory and Its Applications. volume 2. Wiley.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and machine learning. Addison-Wesley.

Hajek, B., 1982. Hitting-time and occupation-time bounds implied by drift analysis with applications. Advanced Applied Probability 14, 502–525.
Happ, E., Johannsen, D., Klein, C., Neumann, F., 2008. Rigorous analyses of fitness-proportional selection for optimizing linear functions, in: Proc. of GECCO '08, ACM Press. pp. 953–960.
Jansen, T., 2013. Analyzing Evolutionary Algorithms. Springer.
Jansen, T., Wegener, I., 2005. Real royal road functions: where crossover provably is essential. Discrete Applied Mathematics 149, 111–125.
Kötzing, T., Sudholt, D., Theile, M., 2011. How crossover helps in pseudo-boolean optimization, in: Proc. of GECCO '11, ACM Press. pp. 989–996.
Lehre, P.K., 2010. Negative drift in populations, in: Proc. of PPSN '10, Part I, Springer. pp. 244–253.
Lehre, P.K., 2011. Fitness-levels for non-elitist populations, in: Proc. of GECCO '11, ACM Press. pp. 2075–2082.
Lehre, P.K., Witt, C., 2012. Black-box search by unbiased variation. Algorithmica 64, 623–642.
Neumann, F., Oliveto, P.S., Rudolph, G., Sudholt, D., 2011. On the effectiveness of crossover for migration in parallel evolutionary algorithms, in: Proc. of GECCO '11, ACM Press. pp. 1587–1594.
Neumann, F., Oliveto, P.S., Witt, C., 2009. Theoretical analysis of fitness-proportional selection: Landscapes and efficiency, in: Proc. of GECCO '09, ACM Press. pp. 835–842.
Neumann, F., Witt, C., 2010. Bioinspired Computation in Combinatorial Optimization – Algorithms and Their Computational Complexity. Springer.
Oliveto, P.S., He, J., Yao, X., 2008. Analysis of population-based evolutionary algorithms for the vertex cover problem, in: Proc. of CEC '08, IEEE Press. pp. 1563–1570.
Oliveto, P.S., Witt, C., 2011. Simplified drift analysis for proving lower bounds in evolutionary computation. Algorithmica 59, 369–386.
Oliveto, P.S., Witt, C., 2012. On the analysis of the simple genetic algorithm, in: Proc. of GECCO '12, ACM Press. pp. 1341–1348. Extended version to appear in Theoretical Computer Science, `http://dx.doi.org/10.1016/j.tcs.2013.06.015`.
Oliveto, P.S., Witt, C., 2013. Improved runtime analysis of the simple genetic algorithm, in: Proc. of GECCO '13, ACM Press. pp. 1621–1628.
Rowe, J.E., Sudholt, D., 2012. The choice of the offspring population size in the (1,λ) EA, in: Proc. of GECCO '12, ACM Press. pp. 1349–1356.
Sudholt, D., 2012. Crossover speeds up building-block assembly, in: Proc. of GECCO '12, ACM Press. pp. 689–702.
Watson, R.A., Jansen, T., 2007. A building-block royal road where crossover is provably essential, in: Proc. of GECCO '07, ACM Press. pp. 1452–1459.
Williams, D., 1991. Probability with Martingales. Cambridge mathematical textbooks, Cambridge University Press.

Appendix A. The Simplified Drift Theorem with Scaling

Theorem 2 is a simplified drift theorem dealing with drift away from the target, which holds in both discrete and continuous search spaces. For its proof, the following lemma will be useful.

Lemma 21. *Let X be a random variable with minimum $x_{\min}$. Moreover, let $f\colon \mathbb{R} \to \mathbb{R}$ be a non-decreasing function and suppose that the expectation $\mathrm{E}[f(X)]$ exists. Then for any $r > 0$*

$$\mathrm{E}[f(X)] \leq \sum_{i=0}^{\infty} f(x_{\min} + (i+1)r)\,\mathrm{Prob}[X \geq x_{\min} + ir].$$

Proof. We denote by p the probability measure from the probability space (Ω, Σ, p) underlying X. Then the expectation is given by a Lebesgue integral, more precisely

$$\mathrm{E}[f(X)] = \int_{\Omega} f(X(\omega))\, p(\mathrm{d}\omega).$$

Since f is non-decreasing and $X \geq x_{\min}$, partial integration yields

$$\begin{aligned}\mathrm{E}[f(X)] &\leq \sum_{i=0}^{\infty} f(x_{\min} + (i+1)r) \int_{\Omega \cap X^{-1}([x_{\min}+ir, x_{\min}+(i+1)r])} p(\mathrm{d}\omega) \\ &\leq \sum_{i=0}^{\infty} f(x_{\min} + (i+1)r)\, \mathrm{Prob}[X \geq x_{\min} + ir].\end{aligned}$$

□

We will use Hajek's following drift theorem to prove our result. Compared to an earlier version of this paper, our presentation of Hajek's drift theorem does not make unnecessary assumptions such as non-negativity of the random variables or Markovian processes. As we are dealing with a stochastic process, we implicitly assume that the random variables X_t, $t \geq 0$, are adapted to the natural filtration $X_0, \ldots, X_t$, $t \geq 0$, though.

We do no longer formulate the theorem using a "potential function" g mapping from some state space to the reals either. Instead, we w. l. o. g. assume the random variables X_t as already obtained by the mapping.

Theorem 4 (Hajek, 1982)**.** *Let X_t, $t \geq 0$, be real-valued random variables describing a stochastic process over some state space. Pick two real numbers $a(\ell)$ and $b(\ell)$ depending on a parameter ℓ such that $a(\ell) < b(\ell)$ holds. Let $T(\ell)$ be the random variable denoting the earliest point in time $t \geq 0$ such that $X_t \leq a(\ell)$ holds. If there are $\lambda(\ell) > 0$ and $p(\ell) > 0$ such that the condition*

$$E\big(e^{-\lambda(\ell)\cdot(X_{t+1}-X_t)} \mid X_0, \ldots, X_t\,;\, a(\ell) < X_t < b(\ell)\big) \;\leq\; 1 - \frac{1}{p(\ell)} \qquad (*)$$

holds for all $t \geq 0$ then for all time bounds $L(\ell) \geq 0$

$$\mathrm{Prob}\big(T(\ell) \leq L(\ell) \mid X_0 \geq b(\ell)\big) \;\leq\; e^{-\lambda(\ell)\cdot(b(\ell)-a(\ell))} \cdot L(\ell) \cdot D(\ell) \cdot p(\ell),$$

where $D(\ell) = \max\big\{1, E\big(e^{-\lambda(\ell)\cdot(X_{t+1}-b(\ell))} \mid X_0, \ldots, X_t\,;\, X_t \geq b(\ell)\big)\big\}$.

Theorem 2 is a simplified version of the scenario underlying the drift theorem. In particular, our formulation does not use moment-generating functions but combines a drift away from the target with a condition on exponentially decaying probabilities for large jumps. It is still relatively general since the exponential decay of probabilities is not required to begin at constant distance, but the distance is allowed to grow with the length of the drift interval.

Proof of Theorem 2. We will apply Theorem 4 for suitable choices of its variables, some of which might depend on the parameter $\ell = b - a$ denoting the length of the interval $[a, b]$. The following argumentation is also inspired by Hajek's work (Hajek, 1982).

Fix $t \geq 0$. For notational convenience, we let $\Delta := (X_{t+1} - X_t \mid X_0, \ldots, X_t \,;\, a < X_t < b)$ and omit to state the filtration $X_0, \ldots, X_t$ hereinafter. To prove Condition $(*)$, it is sufficient to identify values $\lambda := \lambda(\ell) > 0$ and $p(\ell) > 0$ such that

$$\mathrm{E}[e^{-\lambda\Delta}] \;\leq\; 1 - \frac{1}{p(\ell)}.$$

Using the series expansion of the exponential function, we get

$$\mathrm{E}[e^{-\lambda\Delta}] \;=\; 1 - \lambda\,\mathrm{E}[\Delta] + \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{k!}\,\mathrm{E}[\Delta^k] \;\leq\; 1 - \lambda\,\mathrm{E}[\Delta] + \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{k!}\,\mathrm{E}[|\Delta|^k].$$

Since all terms of the last sum are positive, we obtain for all $\gamma \geq \lambda$

$$\begin{aligned}\mathrm{E}[e^{-\lambda\Delta}] \;&\leq\; 1 - \lambda\,\mathrm{E}[\Delta] + \frac{\lambda^2}{\gamma^2} \sum_{k=2}^{\infty} \frac{\gamma^k}{k!}\,\mathrm{E}[|\Delta|^k] \\ &\leq\; 1 - \lambda\,\mathrm{E}[\Delta] + \frac{\lambda^2}{\gamma^2} \sum_{k=0}^{\infty} \frac{\gamma^k}{k!}\,\mathrm{E}[|\Delta|^k] \leq 1 - \lambda\varepsilon + \lambda^2 \underbrace{\frac{\mathrm{E}[e^{\gamma|\Delta|}]}{\gamma^2}}_{=:C(\gamma)},\end{aligned}$$

where the last inequality uses the first condition of the theorem, i. e., the bound on the drift.

Given any $\gamma > 0$, choosing $\lambda := \min\{\gamma, \varepsilon/(2C(\gamma))\}$ results in

$$\mathrm{E}[e^{-\lambda\Delta}] \;\leq\; 1 - \lambda\varepsilon + \lambda \cdot \frac{\varepsilon}{2C(\gamma)} \cdot C(\gamma) \;=\; 1 - \frac{\lambda\varepsilon}{2} \;=\; 1 - \frac{1}{p(\ell)}$$

with $p(\ell) := 2/(\lambda\varepsilon)$.

The aim is now to choose γ in such a way that $\mathrm{E}[e^{\gamma|\Delta|}]$ is bounded from above by a constant. Using Lemma 21 with $f(x) := e^{\gamma x}$ and $x_{\min} := 0$, we get

$$\mathrm{E}[e^{\gamma|\Delta|}] \;\leq\; \sum_{j=0}^{\infty} e^{\gamma(j+1)r}\,\mathrm{Prob}[|\Delta| \geq jr] \;\leq\; \sum_{j=0}^{\infty} e^{\gamma(j+1)r} e^{-j}$$

where the inequality uses the second condition of the theorem.

Choosing $\gamma := 1/(2r)$ yields

$$\mathrm{E}[e^{\gamma|\Delta|}] \;\leq\; \sum_{j=0}^{\infty} e^{(j+1)/2 - j} \;=\; e^{1/2} \sum_{j=0}^{\infty} e^{-j/2} \;=\; e^{1/2} \frac{1}{1 - e^{-1/2}} \;\leq\; 4.2.$$

Hence $C(\gamma) \leq 4.2 \cdot (2r)^2 \leq 17r^2$. By our choice of λ, we have $\lambda \geq \varepsilon/(2C(\gamma)) \geq \varepsilon/(34r^2)$. Since $p(\ell) = 2/(\lambda\varepsilon)$, we know $p(\ell) = O(r^2/\varepsilon^2)$. Condition $(*)$ of Theorem 4 has been established along with these bounds on $p(\ell)$ and $\lambda = \lambda(\ell)$.

To bound the probability of a success within $L(\ell)$ steps, we still need a bound on $D(\ell) = \max\{1, \mathrm{E}[e^{-\lambda(X_{t+1}-b)} \mid X_t \geq b]\}$. If 1 does not maximize the expression then

$$\begin{aligned} D(\ell) \;&=\; \mathrm{E}[e^{-\lambda(X_{t+1}-b)} \mid X_t \geq b] \;\leq\; \mathrm{E}[e^{-\lambda(X_{t+1}-X_t)} \mid X_t \geq b] \\ &\leq\; \mathrm{E}[e^{\lambda|X_{t+1}-X_t|} \mid X_t \geq b] \;\leq\; \mathrm{E}[e^{\gamma|X_{t+1}-X_t|} \mid X_t \geq b], \end{aligned}$$

where the first inequality follows from $X_t \geq b$ and the second one from $\gamma \geq \lambda$. The last term can be bounded as in the above calculation leading to $\mathrm{E}[e^{\gamma|\Delta|}] = O(1)$ since that estimation uses only the second condition, which holds conditional on $X_t > a$. Hence, in any case $D(\ell) = O(1)$. Altogether, we have

$$e^{-\lambda(\ell)\cdot\ell} \cdot D(\ell) \cdot p(\ell) \;\leq\; e^{-\ell\varepsilon/(34r^2)} \cdot O(r^2/\varepsilon^2) \;=\; e^{-\ell\varepsilon/(34r^2)+2\log(r/\varepsilon)+O(1)}$$

By the third condition, we have $r^2 \leq \varepsilon\ell/(132\log(r/\varepsilon))$. Therefore,

$$\frac{1}{2} \cdot \frac{\varepsilon\ell}{34r^2} \geq 2\log(r/\varepsilon),$$

which finally means that

$$e^{-\lambda(\ell)\cdot\ell} \cdot D(\ell) \cdot p(\ell) \;\leq\; e^{-\ell\varepsilon/(68r^2)+O(1)}$$

Choosing $L(\ell) = e^{\ell\varepsilon/(132r^2)}$, Theorem 4 yields

$$\mathrm{Prob}[T(\ell) \leq L(\ell)] \;\leq\; L(\ell) \cdot e^{-\varepsilon\ell/(68r^2)+O(1)} \;=\; O(e^{-\varepsilon\ell/(132r^2)}),$$

which proves the theorem. □

The scaling factor $r(\ell)$, which is only relevant for the second condition of the theorem, can be bounded from above by a constant in many applications to evolutionary algorithms.

Appendix B. Non-monotone Variable Drift

Proof of Theorem 1. We define $g(z) := \frac{z_{\min}}{h(z_{\min})} + \int_{z_{\min}}^{z} \frac{1}{h(x)}\,\mathrm{d}x$ for $z \geq z_{\min}$ and $g(0) := 0$. Note that g is invertible and that the first hitting time of 0 for the Z_t-process is the same as the first hitting time of 0 for the mapped process $g(Z_t)$, $t \geq 0$. Assuming $Z_t \geq z_{\min}$, we compute the drift of the mapped process

$$\begin{aligned} &\mathrm{E}[g(Z_t) - g(Z_{t+1}) \mid \mathcal{F}_t] \;=\; \mathrm{E}\left[\int_{Z_{t+1}}^{Z_t} \frac{1}{h(x)}\,\mathrm{d}x \;\middle|\; \mathcal{F}_t\right] \\ &= \mathrm{E}\left[\int_{Z_{t+1}}^{Z_t} \frac{1}{h(x)}\,\mathrm{d}x \cdot \mathbb{1}\{Z_{t+1} < Z_t\} \;\middle|\; \mathcal{F}_t\right] - \mathrm{E}\left[\int_{Z_t}^{Z_{t+1}} \frac{1}{h(x)}\,\mathrm{d}x \cdot \mathbb{1}\{Z_{t+1} > Z_t\} \;\middle|\; \mathcal{F}_t\right]. \end{aligned}$$

Item (4) from the prerequisites yields $h(z) \leq ch(Z_t)$ if $Z_t - d(Z_t) \leq z < Z_t$ and $h(z) \geq h(Z_t)/c$ if $Z_t < z \leq Z_t + d(Z_t)$. Using this and $|Z_t - Z_{t+1}| \leq d(Z_t)$, the drift can be further bounded by

$$\mathrm{E}\left[\int_{Z_{t+1}}^{Z_t} \frac{1}{ch(Z_t)}\,\mathrm{d}x \cdot \mathbb{1}\{Z_{t+1} < Z_t\} \;\middle|\; \mathcal{F}_t\right] - \mathrm{E}\left[\int_{Z_t}^{Z_{t+1}} \frac{c}{h(Z_t)}\,\mathrm{d}x \cdot \mathbb{1}\{Z_{t+1} > Z_t\} \;\middle|\; \mathcal{F}_t\right]$$

$$\geq \mathrm{E}\left[\int_{Z_{t+1}}^{Z_t} \frac{1}{2ch(Z_t)}\,\mathrm{d}x \cdot \mathbb{1}\{Z_{t+1} < Z_t\} \;\middle|\; \mathcal{F}_t\right] = \frac{\mathrm{E}[(Z_t - Z_{t+1} \mid \mathcal{F}_t) \cdot \mathbb{1}\{Z_{t+1} < Z_t\}]}{2ch(Z_t)}$$
$$\geq \frac{h(Z_t)}{2ch(Z_t)} = \frac{1}{2c},$$

where the first inquality used the Item (2) from the prerequisites and the last one Item (1). An application of the classical additive drift theorem with drift $1/(2c)$ and initial distance $g(Z_0)$ completes the proof. □